Qualifications and
Curriculum Authority

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

Alf Massey, Sylvia Green, Trevor Dexter and Lisa Hamnett

Research and Evaluation Division

University of Cambridge Local Examinations Syndicate

Comparability of national tests over time:

key stage test standards between 1996 and 2001

Final report to QCA of the Comparability Over Time Project

Alf Massey, Sylvia Green, Trevor Dexter and Lisa Hamnett
Research and Evaluation Division, University of Cambridge Local Examinations Syndicate

Contents

Comparability of national tests over time:
key stage test standards between 1996 and 2001

Alf Massey, Sylvia Green, Trevor Dexter and Lisa Hamnett
Research & Evaluation Division, University of Cambridge Local Examinations Syndicate

1 Introduction, literature review & research outline

The Project's origin and brief

The Project was instigated to investigate the equivalence of standards set in national tests over a period of several years. A contract for this Qualifications and Curriculum Authority (QCA) funded work was awarded to the Research & Evaluation Division of the University of Cambridge Local Examinations Syndicate (RED, UCLES) in April 1999. Work was concluded at the end of 2001; with the report submitted in January 2002.

Two strands of quantitative research were commissioned initially. The first involved experimental comparisons - beginning with KS1 Reading Comprehension, KS2 English and KS2 Mathematics tests set in 1996 and 1999. These experimental comparisons were subsequently extended to encompass all key stages and subjects where national tests were in use. Phase 2 included comparisons of the KS1 Mathematics, KS2 English (a replication) and KS3 Mathematics tests set in 1996 and 2000 and phase 3 compared the KS2 Science, KS3 English and KS3 Science tests set in 1996 and 2001. The second strand was a search for information on changes in achievement in schools from an alternative source - standardised testing programme databases held by Local Education Authorities (LEAs).

As the Project's work developed, three strands of 'qualitative' research were also commissioned. One asked children to consider selected features in test materials, as they have evolved in recent years, to discover which they felt were important and how they affected their preferences for the tasks involved. Another asked teachers to compare sets of scripts 'representing' key mark points from two versions of a national test - one past and one current - to see if their judgements concurred with other findings. The third qualitative strand allowed the Project to observe the current procedures for setting level thresholds for national tests in action - the key to national test standards. This enhanced our understanding of the ways in which the operational development cycle sought to maintain standards over time; providing a basis for a brief critical evaluation and focussing the project's reflections on the policy implications of the other data gathered.

The political dimension to comparability of standards over time

Comparability of standards over time is fundamental to the credibility of any examination or assessment scheme setting 'equivalent' tests in different years. The evidence provided by test results across the years is often central

to the discussion of key policy issues. For instance, the Department for Education and Employment's recent Green Paper, 'Schools - building on success' (DfEE, 2001a) draws on such evidence to assert that

- 'more children leave primary school able to read and write well. Seventy-five per cent of children achieved Level 4 in 2000 compared to just 57 per cent in 1996.....'
- 'more children leave primary school numerate. Seventy-two per cent achieved Level 4 in 2000 compared to 54 per cent in 1996....'

A main strand of the 'National Literacy Strategy' (Beard, 1998) set a target of raising the proportion of 11 year olds reaching the standard expected for their age in KS2 English tests from 57% in 1996 to 80% in 2002. But even this ambitious target was not enough and in March 2001 the DfEE announced plans for new, tougher and more extensive targets (in all subjects tested) for 2004 for pupils reaching the end of both KS2 (aged 11 years) and KS3 (aged 14 years), together with longer range targets for 2007. Target setting, in terms of national test results, driven by local targets and comparisons between schools and Local Education Authorities, is high on the policy agenda and the creation of a target setting culture in schools (DfEE, 1997) is seen as a key to levering up standards of achievement. The importance of rigorous consistency, year on year, in standard setting for national tests has been recognised (LTF, 1997) if test results are to serve as an effective yardstick for measuring improvement. If test standards slip, apparent progress will be only an illusion.

Comparability can clearly be a matter of contention, especially for high stakes tests like those we have investigated. Indeed, shortly after this project began, the press observed that 1999's KS2 national test level thresholds were lower than 1998's and suggested that 'standards' were being eased so that national targets for the achievements of 11 year olds would be met. The extensive press speculation on the basis of such one dimensional evidence was clearly either naively misinformed or mischievous, but in view of the public interest, the Secretary of State for Education and Employment instigated an immediate enquiry by an Independent Panel. The Panel was asked to consider the arrangements for setting and maintaining standards in the tests. It reported swiftly (Rose et al, 1999), indicating that it was confident that the press reports were unfounded; that procedures in 1999 were similar to those in earlier years; and that there was no reason to question the 1999 test results. The panel's report described the procedures for test development and standards setting in some detail. It recognised that 'there will always remain a degree of subjectivity', especially in English tests - where most criticisms had arisen, and suggested minor improvements for the future. Amongst these, the Panel recognised the role of this Project, which formed part of the ongoing management of the testing programme, and advocated further similar work.

To avoid further brouhaha about the deliberate manipulation of test standards, it is important we declare unambiguously that whatever conclusions are reached by this Project, there can be no possible conflict with the Rose Panel's conclusions. The Panel did not have the benefit of evidence like ours.

It could only determine whether the proper procedures had been followed in setting and marking the 1999 tests and was in no doubt that they had. We are not trying to second guess the Rose Panel's conclusions regarding procedural matters, but are instead investigating the underlying issue they were unable to address; namely, whether or not the use of proper procedures over an extended period had successfully maintained test standards.

The research literature

Better results v declining standards

The improvements in the literacy and numeracy of children leaving primary schools in England between 1996 and 2000 quoted above refer to the results in KS2 mathematics and English tests. These fairly dramatic shifts are not untypical of the results in national tests in other subjects and key stages since they were introduced. In all subjects at all key stages the trend in test results is upwards (DfEE, 2001b). The same could be said of public examinations in England and Wales at ages 16 and 18 (OFSTED & SCAA, 1996) and the issue is not confined to the UK - for instance in the USA the phenomenom akin to this is described as grade inflation (Ziomek and Svec, 1995).

But other research has suggested that standards of achievement in basic skills may be remarkably stable. Brooks, Foxman and Gorman (1995) surveyed the literature concerning standards in Literacy and Numeracy in the UK between 1948 and 1994 and concluded that reading standards at 11 had changed little since 1945, except for small rises around 1950 and in the 1980s. Among 6-8 year olds however, standards fell slightly in the late 1980s. In writing performance there were no changes during the 1980s. In number skills they detected a fall in the attainment of 10/11 year olds nationally between 1982 and 1987, although achievement in geometry, statistics and measures rose in compensation. Galton (1998) was even more pessimistic and suggested that his comparisons of children in years 4, 5 and 6 in 1977 and 1997 showed decline, rather than stagnation, in mathematics, reading and language skills.

Against this unpromising background, it is easy to see why the upward trend in results of national tests is questioned. Critics often reason that there is no reason to believe that the current cohort of children is more able than their immediate predecessors and doubt that improvements in the efficiency of schooling can bring about gains on the scale implied by improving results, in so short a time scale. The logic of the argument that intelligence has a substantial hereditary basis and that large environmentally induced gains are unlikely in the short-term is tempting, if inherently pessimistic. But its foundations would be weak if it could be shown that recent generations were more able than their forbears. A credible, though not yet well known, case has been made for this. Dickens and Flynn (2001) reviewed an extensive international body of research and presented a formal model of the process determining IQ in which IQs are affected by both environment and genes, where environments are matched to IQs. This feature of the model allows very large environmental 'multiplier' effects, of an order which might explain

the large post-1950 gains observed on a variety of different tests of intelligence in over 20 nations. The pessimists may yet be proved wrong.

Research into national tests

The research literature contains little that has directly addressed the stability of standards in national tests over time. ‘Standards’ do attract considerable attention, as do the results of national tests, but too often debate amounts to little more than a deliberate ‘spin’; taking advantage of information to make a partisan point (Fox, 1998). Serious research on the new national assessment system has tended to concentrate on its impact on teaching and learning. For instance Clarke (1997) pointed out how testing at Key Stages 1 and 2 might reduce the effort available for teaching and learning and redirect teachers' attention towards the goals which featured in the tests. Brown et al (1996 & 1997) investigated schools’ reactions in some detail and they also saw Key Stage 2 tests as a ‘curricular magnet’. They were able to point to ways in which schools had changed their practice, both in adapting teaching to the tests and providing practice opportunities and in the ways in which some schools used the test results. Whilst their evaluation was mixed, their work suggested that schools were likely to make increasing use of information from the tests and that many teachers believed that tests had a valid role, at least in the short term. Similarly, Preece and Skinner's (1999) evaluation of KS3 Science tests focussed mainly on their impact on teaching and learning.

In the initial years of national tests in England and Wales the test construction and standard setting approaches tried were radically different from those used in previous large scale testing programmes, on either side of the Atlantic. The basic building blocks of the '1991 edition' of the national curriculum were a multitude (Science had 176 for instance) of 'statements of attainment' (SoAs) purporting to describe target behaviours. SoAs were grouped within a matrix of Attainment Targets (ATs) (Science had four of these) and the ten national curriculum 'levels'. Sizmur and Sainsbury (1997) describe how this atomistic method of defining national curriculum levels proved overly cumbersome for teachers and the same proved true for testing.

Massey (1995) summarised the speedy evolution of national tests: from narrow SoA based criterion referencing in 1992, via a system crossing criterion-referencing with domain-referencing in sub-tests targeted at each AT at each level in 1993, to tests reporting an overall 'test level' through level 'cut-scores' on the total score scale in 1994. He described the use of a 'criterion-related' test development model designed to produce a consistent series of tests and cut-scores linked to 'level descriptions' introduced after 1995 to replace SoAs (and by then already viewed as a descriptive set). But even such weak forms of criterion-referencing in the national test development process may have petered out, with no further work in this vein being reported. This at least has the advantage of allowing post-hoc standard-setting and equating (and investigation of comparability) to make use of a wide range of established methods. National testing in the UK now only displays weak features of criterion-relatedness (Sainsbury and Sizmur, 1998) as level descriptions have little direct bearing on standard setting arrangements.

Comparability over time

Comparability over time is not at all easy to investigate (Goldstein, 1983; Newton, 1997). Even if all other things stayed the same from one year to another, the difficulty of tests set in different years will change, as changes in content and various features of the questions alter the likelihood of children answering correctly. Thus the scores obtained by 'equivalent' children in different years will vary.

Clearly we cannot simply expect that a mark of x in one year will mean the same thing as a mark of x in another. It is for this reason that standard setting in national tests requires an annual judgemental exercise to set cut-scores for each level which reflect achievement equivalent to previous years. In the course of this, systematic professional judgements about the difficulty of the tests set are considered alongside equating study data from the test development process and from national samples of children taking the operational tests (Rose et al, op.cit.).

Shifting sands

Other things do not necessarily stay the same of course, adding further complications, especially over the longer term. Policy decisions change the curriculum's contents and these and other initiatives affect what schools teach. Patrick (1996) described the many ways in which the curriculum and assessments have changed over the years and pointed out how they made meaningful comparisons more difficult. The pace of change has not lessened recently. For instance the national curriculum for England was first introduced in 1989 and subsequently extensively revised in 1993. Within this period we have seen a new OFSTED schools' inspection regime; the introduction of wide scale use of assessment data for value added analyses and benchmarking, so that schools can compare their performance against others; publication of league tables of schools, based on national assessments and public examination results and the introduction of target setting for schools' results; the specification of Desirable Learning Outcomes for pre-school education in 1996; availability of optional tests and assessment units for years 3,4 and 5 in 1997/8; new Baseline Assessments on entry to schools from 1998; and the advent of the National Literacy and Numeracy Strategies. Even the forms of the national tests themselves have been revised quite extensively over the years, affecting the ways in which children will be prepared for them.

In addition to such changes within the educational system we must also recognise the potential impact of cultural change. For instance Massey and Elliott (1996) described some of the changes between 1980 and 1994 in the style of children's writing and the errors they made and showed how these reflected changes in public values.

A variety of methodological options for comparing standards over time have been explored. Some are however more useful than others.

Stratified comparisons of distributions
The examining boards responsible for school examinations in England and Wales have made use of the systematic differences in success rates between different types of schools and colleges to improve upon simply comparing the distributions of grades awarded from one year to another (Massey (1978); Massey & Newbould (1978)). On the assumption that the calibre and achievement of candidates from each type remained stable, results from different types over a period were used to create expected distributions, which were compared with actual outcomes. This may have some value as an inexpensive screening device and helps to monitor the possible effects of changing entry patterns in examinations. But the assumptions made are untenable - ruling out any real rise or fall in performance standards if teaching and learning improved over time.

Judgemental comparisons of the equivalence of level of demand
The juxtaposition of the question papers and tasks from different periods and the assertion that one (usually the later) is less demanding than the other is common journalistic fodder and can take more scholarly forms. For instance Hilton (2001) categorised the questions from the KS2 Reading tests set in 1998, 1999 and 2000 on a common basis and asserted that the 2000 test contained fewer questions requiring higher-order reading skills. From this she concluded that the 2000 test was 'easier' and questioned the rising trend in results. But the skills demanded are only one of many factors which will affect the difficulty of test items. Small features can have great effects (Pollitt et al (1985); Pollitt and Ahmed (1999); Ahmed and Pollitt (1999 & 2001)) and question or test difficulty cannot be predicted safely by desk research. For instance we shall show empirically, later, that the 2000 KS2 Reading test (studied by Hilton) was more difficult than the 1996 version. But one certainly should not conclude from this data that test standards had been raised (not lowered as Hilton concluded), let alone from an analysis of the level of demand. It is the combination of test difficulty (which controls the distribution of marks scored) and the level thresholds adopted which govern standards. Hilton's approach ignored the role of thresholds too.

Though concerns about 'demands' may be valid and important, they are insufficient to support inferences about test 'difficulty' or 'standards'. Judgements about the difficulty of test items are famously unreliable. For instance Good and Cresswell (1988) demonstrated that even very experienced examiners found it very hard to construct questions of specified difficulty without empirical performance data.

Judgemental comparisons of the quality of work
Cresswell (2000) has also shown how panels of examiners responsible for grade thresholds in public examinations can identify the direction of changes in the difficulty of examination papers from one year to the next, but find it difficult to determine the extent of such changes and tend to under-estimate them substantially.

Such lack of precision notwithstanding, expert judgements are a necessity if the quality of work represented in archived scripts is to be compared with that

of candidates from the modern era. Christie and Forrest's (1981) sophisticated attempt to investigate AL grading standards over a ten year gap recognised the importance of investigating the variations in criteria implicit in examinations from different periods, as well as trying to evaluate the relative merits of candidates' work. But judgements are never value free (Cresswell, 2000), and changes in what is valued are at the heart of curricular change over the years. The examiners and values from the past are inevitably under-represented in such comparisons. Similarly, examiners' judgements have more recently been employed to compare GCE AL Mathematics grading standards between scripts from 1985 and 1996 (Bell, Bramley & Raikes, 1998; Bramley, Bell & Pollitt, 1998) using paired comparisons. But whilst human judgement has the flexibility to take changing conditions and criteria into account, albeit subject to the arguable effects of changing values, it inevitably finds it difficult to detect relatively small differences with any confidence, especially if comparisons are complex and/or evidence is slim. For instance, the 2001 report from the quasi-operational rolling programme of 'five yearly standards reports' (produced by QCA, as the regulatory body) on GCE and GCSE standards, describes how independent subject and assessment experts reviewed syllabus materials and archived examples of candidates' work in 7 GCSE subject examinations and 7 GCE subject examinations against their contemporary equivalents (QCA, 2001). These involved comparisons over time intervals ranging from three to twenty years. In all 14 cases the reviewers concluded that standards had been maintained. But will such imprecise methods do where national tests are the key indicator (Earl et al, 2001) to the success of major policy initiatives? The lack of power in judgemental approaches is a serious deficiency in a policy oriented research field, where shifts of a percentage point or two in national distributions have been hailed as success or disaster in the media.

Analytical comparisons of the quality of work

Analytical comparison of the quality of work from different eras is also possible. But more detailed analysis of features of children's work does not have the flexibility of judgements to compensate for differences in the tasks set and the value systems in play. This approach is rare, perhaps because of the lack of suitable archived work from substantial numbers of candidates. An example is provided by Massey & Elliott (op.cit.), who compared the measurable features of single sentences sampled from the work of large numbers of candidates with equivalent grades in English examinations in 1980 and 1993/4. Whilst able to conclude that in 1980, amongst other things, children used more adventurous vocabulary and sentence structures, were as likely to be grammatically accurate and made less than half the number of spelling mistakes, they were unable to conclude that examination grading standards had declined. The examinations in question were very different, with children being asked to tackle a much wider range of tasks in 1994, and changes in writing were related to changing social and linguistic conventions over this period.

Common reference tests

Another possibility is the use of a separate common reference test to collect data about the abilities of samples of children taking tests in different years,

so that the 'levels' achieved by equivalent children (i.e. those obtaining the same scores on the reference test) each year can then be compared.

Recently Fitz-Gibbon and Vincent (1994) used an aptitude test to investigate grading standards in AL mathematics over a seven-year period and Coe (1999) used the same data source to make comparisons between grades in six A level subjects over a ten-year period. But the capacity of this approach to provide definitive answers to questions of comparability has been in doubt for many years. A general ability reference test was tried extensively in comparing CSE and GCE grading standards in England and Wales between 1966 and 1976 (Willmott, 1997 &1980) but the approach was abandoned because it was apparent that the relationship between the reference test and achievement in examinations may not be constant across sub-groups of candidates or over time. This is an implicit assumption of the method, along with the view that the factors involved in teaching and learning are also constant. The latter assumption is also clearly untenable. Successive cohorts are likely to obtain similar scores on any measure of general ability, whilst interventions to improve educational practice seek to raise achievement scores. By definition, if in later years schools and/or changes in the educational system succeed in improving educational outcomes, the measured achievements of pupils of any given level of ability would be higher than their forerunners'. Why should this be interpreted as a lowering of the test's standards for awarding grades/levels, rather than rising standards of performance by teachers and pupils? Public discussion of these issues is hampered when several such contradictory definitions of standards are confused in common usage (see Massey, 1994).

Subject specific reference tests prove problematic too. Newbould and Massey (1979) came to similar conclusions about the variations in relevance and bias evident even with the common elements found within some examinations, which also inhibited firm conclusions about the equivalence of standards.

We must conclude that reference tests are unlikely to prove definitive in research of this nature. But they do remain useful and relatively inexpensive tools to alert the managers of assessment programmes to potential problems. This strategy has indeed been used in this spirit to monitor equivalence across the different KS3 national testing programmes (Massey, 1998) in England, Wales and Northern Ireland.

There may also be opportunities to apply this conceptual and analytic approach to pre-existing sources of concurrent data, independent of the tests for which standards are being compared. For instance Hurry et al (1996) suggest that some schools and/or Local Education Authorities have continued to employ the same standardised tests over an extended period. Such data might allow comparisons of the relationship between children's standardised test scores and their national test levels, between successive cohorts. But to make use of such 'found' data one must be sure that assessments have been made on the same basis over the period concerned. For instance, any doubts about the stability of standards set in KS2 national tests in successive years would invalidate the use of KS2 results as the 'common' foundation for

comparisons between KS3 tests set in different years. If the relevance or validity of found measures, such as standardised tests, were to change over time, children of similar ability might obtain systematically different scores, so undermining the methodology. Small-scale studies of this type have been reported. Davies and Brember (1995) investigated the relationship between standardised test scores and KS1 Mathematics levels in 1991 and 1992 and found that the mean test scores for pupils at each level had fallen. They argued that this cast doubt on rises in standards as measured by national assessments. Davies (1999b) reported similar falls in mean standardised reading test scores for children achieving the same KS2 Mathematics levels in 1995 and 1996, leading to similar conclusions.

Experimental comparisons
Direct experimental comparisons have for long been seen as a means of equating tests. For instance Angoff (1971) provided a series of experimental designs for this purpose. 'Equivalent' groups of children are asked to take different versions of the tests and, if standards are in line, are expected to obtain similar results. Although a recent small scale experiment in one primary school (Brown, 1999) suggested that children (who took both) obtained higher levels on the 1999 KS2 English tests than the 1998 version, this approach has not been used to investigate comparability over time in the UK on a large scale. There may be various reasons for this. Chief amongst these are probably high costs and the difficulty of identifying a suitable sample for whom both a past and current test being compared are equally appropriate, in terms of the curriculum followed and conditions of motivation. Massey (1997) reported work from a national test development programme relevant to both of these issues. Curricular appropriateness was shown to affect the outcome of equatings between test levels achieved by children from different key stages who were asked to take the same tests, illustrating the ways in which equatings or evidence regarding comparability produced using data from a group from one curricular tradition are unlikely to remain valid for others. An example of the quasi-experimental comparisons which form part of the operational procedures for equating national tests was also described, with a final trial of each test taking place close to the operational testing of the children taking part. Operational test data is then used in an equating between the current and future versions of the test. But the fact that children's motivation levels may vary between their operational test and a trial has always been recognised as a problem and the effect of this on equatings is unpredictable, limiting their precision.

Project outline

The Project was fortunate in being able to employ more than one approach. Two chief empirical research strategies were used to investigate the constancy of national test standards over the period 1996-2001, each based on different operational definitions of equivalence. We were also fortunate in being able to support these with three strands of qualitative research - one designed to follow up empirical work by seeing if teachers' judgements about the quality of children's work would support the findings of empirical

comparisons; another to investigate children's reactions to test materials and the ways in which the accessibility of these may be changing; and a third which considered the ways in which national test standards are set and helped us to reflect on our findings and produce recommendations to contribute to debate about arrangements for national tests at KS1, KS2 and KS3 in future years.

Quantitative Strands

Direct experimental comparisons
This Project employed a strategy previously untried in the context of large scale assessment in the UK - direct experimental comparisons. Operational equivalence here assumes that groups of children assigned at random (and hence in this sense 'matched') to different versions of the test should obtain similar results if standards are in line. This seemingly obvious (but rarely used) approach was feasible here as we were able to overcome the chief obstacles: high costs and unequal curricular biases and motivational conditions for the tests compared.

The relatively high costs incurred are justifiable in the context of a large-scale national testing programme where comparability over time is a crucial issue. It also proved feasible to identify suitable samples, largely because the comparisons attempted here involve time-spans of three to five years, namely 1996 to 1999, or 2000, or 2001; over which the national curriculum and tests had remained relatively stable. Over this medium term cultural or curricular changes should not invalidate comparisons; though we took this as something to be investigated too and explored the possibility that the two versions of the tests compared in each case could differ in ways which might bias the comparisons. Experimental comparisons over longer periods where curricular change can be seen to affect the nature of the tests concerned would be much more likely to prove problematic. Teachers might be unwilling to submit their pupils to inappropriate tests and the results would anyway be contaminated.

In this Project the tests were set to children from Northern Ireland, where a similar (but not identical) national curriculum to England's is in operation. These children were able to take both versions compared under similar 'experimental conditions', including, critically, their level of motivation to perform well. Asking children in England to do tests from a previous year as well as their 'live' tests inevitably creates a difference in motivation between forms for different 'years'. Experiments involving two previous versions would avoid this but the latest versions could not be used - a serious deficiency given the immediate policy relevance of this research. There would also be a high risk that participants had been exposed to the materials in the course of teaching or practice tests. But none of the tests involved 'high stakes' for the children in Northern Ireland and they would not have encountered them previously. Although this ensured equivalent conditions we recognised that the validity of testing the null hypotheses with a sample of pupils from

Northern Ireland must also be investigated, as their schooling might vary in ways that could bias the results.

A full description of this strand of the Project's work is given in Section 2 below.

Evidence from standardised testing in schools
National tests are not the only large-scale tests of achievement used in schools. Some Local Education Authorities encourage their schools to participate in standardised testing programmes for various purposes and these programmes may hold alternative evidence suggesting a rise or fall in standards of achievement over the years.

The second strand within the Project seeks to cross-validate outcomes from the experimental strategy above by locating and considering evidence concerning changes in the relationships between the outcomes of standardised testing programmes in Local Education Authorities and national test results, from 1996 onwards.

The operational definition of equivalence inherent in this approach is that standards would be the same if, on average, children from successive cohorts through the years, who have similar scores on a given standardised test, obtain equivalent results in national tests. In essence this uses the 'available' standardised test data from children tested in different years as a common reference test to monitor standards in the (different versions of) national tests set each year.

We have of course already indicated that this methodological approach has been questioned and cannot in itself provide a definitive answer to the question 'have standards of national tests changed?' But whilst not providing a definitive answer it may still make a useful contribution. It looks at the data from individual local authority testing programmes in a new light, compares them and sets them in a wider context. This might illuminate some concerns and shed light on issues the experimental comparisons cannot address, such as the progress of standards across the intervening years between 1996 and 2001.

This strand of the Project is reported in Section 5 below.

Qualitative Strands

Teachers' judgements of 1996 and 1999 KS2 English scripts
Would teachers looking at children's scripts agree with the conclusions from our experimental comparisons? We attempted to confirm our findings in KS2 English by mounting a small-scale judgemental study. The limitations of judgemental methods have been discussed above and are not disputed. But it provided an opportunity to show a group of teachers some relevant evidence, to see what they made of it and to hear what they said about both the evidence and the issues involved. This exercise is reported in Section 3.

Children's perceptions of selected features of tests materials
National Tests themselves have not (so far) stood still. QCA and their development agencies have introduced modifications designed to improve the tests even during this relatively stable period since 1996. How do children see such changes? Do they notice them at all? Do they like them? Might they make it easier for children to show what they can do?

We mounted an initial small-scale qualitative investigation of such issues, using semi-structured interviews with samples of children to explore features in the 1996 and 1999 KS2 English tests. The voicing of children's views provided some interesting and novel insights and we were asked to mount similar investigations in the other subjects and key stages, so that we could consider their implications for our interpretation of the empirical evidence from the other strands of work described above.

This research is reported in Section 4.

A small-scale evaluation of level setting procedures
Evaluation is perhaps too grand a term for this strand of work. The Project was asked to observe at a selection of the key operational meetings etc. leading to the setting of level thresholds at KS2 and KS3 in 2000. The intention was to enable an 'outsider', albeit someone with knowledge and experience of the national testing system and UK schools examinations, to produce a critical review of the procedures which have evolved through the early years of this testing programme. An interim report to the Qualifications and Curriculum Authority was produced, but was largely a personal view and was neither intended nor suitable for publication. However this report does draw upon the insights gained where relevant.

2.1 Experimental Comparisons: Methodological Issues

The experimental comparisons: what, when and where

Over the Project's three-year life, experimental comparisons were made at all key stages and in all subjects where national tests are in use.

In all cases the comparisons involved the 1996 versions of the tests, this being the earliest point in time when national tests across the range of subjects and key stages were seen as 'settled', following a period of initial development and revisions of both the curriculum and the format of the tests. The 'later' versions compared with the 1996 tests were the most recent available. In phases 1, 2 and 3 of the Project respectively, these were the tests from 1999, 2000 and 2001. The full set of experimental comparisons made are listed below.

Phase 1
- KS1 Level 2 Reading Comprehension test: 1996 v 1999
- KS2 English: 1996 v 1999
- KS2 Mathematics: 1996 v 1999

Phase 2
- KS1 Mathematics: 1996 v 2000
- KS2 English: 1996 v 2000[1]
- KS3 Mathematics: 1996 v 2000

Phase 3
- KS2 Science: 1996 v 2001
- KS3 English: 1996 v 2001
- KS3 Science: 1996 v 2001

The elapsed time between the versions of the tests involved in these comparisons thus varied from three to five years, across a period of relative curricular stability - at least by comparison with the early 1990s. Even so, there were a few notable curricular changes and/or alterations to the structure of some tests; the effects of which will be discussed when the comparisons concerned are reported. But it seemed likely that it would be feasible to ask children to sit versions of national tests spanning this period without the earlier version being so outmoded as to make comparisons unfair. If comparisons are to be unbiased, it is vital that contemporary teaching and learning are relevant to both versions of the test.

Such medium term experimental comparisons are perhaps the best we can expect. With any shorter interval there could have been too little time and opportunity for standards to drift for discernible effects to have come into being. But the curriculum will not stand still. With any longer gap we would be more likely to encounter changes to the curricular and assessment regime which will threaten the validity of experimental comparisons - like the review of current arrangements for national assessment in England expected to lead to

[1] A replication, included to verify the methodology and because of the attention attracted by this subject/key stage.

changes in the system for 2003. Culture and practice in schools can also, like fashion, move on surprisingly quickly.

Finding suitable schools/children to take part in fair experimental contrasts between different versions of national tests is less straightforward than it might at first seem, especially if it is hoped to include the latest available versions of the tests. By definition all children in England of the appropriate age will take the latest versions operationally. Their motivation to perform well must be higher for these than for another test, taken only for research purposes. Past tests are also used for practice in schools, to prepare for operational testing, and it would be difficult to find samples without prior experience of the previous (or any given earlier) version in schools in England. The Project instead used schools in Northern Ireland (NI), where the curriculum and testing regime is very similar to England's, but not identical, and where any two versions of these national tests would both be 'low stakes' assessments, with children having no reason (extrinsic to the tests themselves) to be more motivated to do well on one version than another.

Checking the validity of the experimental comparisons

To be confident that experimental comparisons using children in Northern Ireland, rather than England, are valid, we need to be assured that there is nothing about test forms from different years, and/or the education of children in Northern Ireland, which might invalidate our comparisons. We would not be concerned if the children involved were likely to score higher or lower on the tests than English children of equivalent ability, as we are not concerned to measure the achievement of NI children or compare it with performance levels in England. Indeed the sample would not enable any such comparisons to be made. Our concern is only about the fairness of the comparisons between test forms from different years. The key question here is simple. Might we have reason to expect NI children to find the 1996 version of a test harder or easier than the 1999 version? The Project pursued this question on several fronts.

Variations between the 1996 and 1999/2000/2001 versions of tests
Project staff reviewed the contents of the versions of the tests being contrasted at each key stage/subject in detail; considering the types of questions and stimuli (such as texts) and the topics and skills tested by each question; and seeking to identify variations between the versions involved. Such differences are not in themselves significant, but become so if the scores obtained by children from England seem likely to be higher or lower than those from Northern Ireland on one version rather than another to which it is being compared. The likelihood of this was considered in light of differences between the two educational systems, which were investigated as described below.

Structural and curricular differences between England and Northern Ireland
Project staff undertook detailed reviews of the official documentation describing the structure of the two school systems and the curriculum (in each subject/key stage where comparisons were made in turn) in England and in

Northern Ireland; with a view to identifying relevant differences, so that the possibility that these might bear unevenly on the versions of the tests being compared (and so bias our experimental findings) could be evaluated.

Teachers' opinions

The desk reviews above could only point to aspects of the curriculum where practice might differ between England and Northern Ireland. The curriculum documents reviewed covered much the same ground, but were not arranged in exactly the same way. How far teachers' classroom practice might be affected by (often small) differences or omissions in wording is a matter of conjecture. It is possible that what is taught varies less - or more - than the documentation might suggest. The Project therefore sought the views of the teachers in the schools where testing was carried out.

Different methods of gathering teachers' views were used in Phase 1 of the Project to Phases 2 and 3. In Phase 1, approximately one month after the tests were administered, project staff visited each school participating in the testing programme and talked to heads/key stage co-ordinators and/or assessment co-ordinators and classroom teachers. Discussion covered general matters, including confidential reports the Project had prepared on each school's performance compared to others and their experience of administering the experimental tests. It also included a series of questions asking if pupils' learning experiences would have prepared them adequately for the two versions of the test or might have left them ill-prepared for particular questions, or led them to be especially well prepared for others. Fieldwork staff used a structured interview schedule to guide discussions and to structure their reporting. In general the teachers were positive about the format, content and presentation of the tests and considered them appropriate for children in Northern Ireland, as well as being well balanced and easy to administer, although some Key Stage 2 teachers wondered if performance might have been affected by poor motivation, following a relatively high assessment load towards the conclusion of KS2. Comments on specific features of the KS1 Reading Comprehension and KS2 English and Mathematics tests administered in Phase 1 are summarised in the reports on each comparison, which follow.

The fieldwork in Phase 1 proved helpful, not least in helping project staff understand some of the nuances of schooling in Northern Ireland. But in Phases 2 and 3 it was felt that the same purposes could be served more efficiently by other means. In Phases 2 and 3, confidential reports for each school were sent by post and telephone support was offered to help schools wishing to have these explained or to discuss them. Teachers' views on the appositeness of the test materials had been sought by means of questionnaires, which had accompanied the test materials. Like the fieldwork interviews, these enquired about the general suitability of the two versions of the test in the Northern Ireland context as well as asking teachers to identify questions where children's preparation might affect their performance and say why. Schools proved very helpful in responding to these and their views are summarised in the reports on each comparison which follow.

Research design

A 'supplemented' equivalent groups design
The Project in essence employed an ‘equivalent groups’ experimental design. Children were randomly assigned to take either the 1996 or the latest available (1999, 2000 or 2001, according to the phase of the Project, as described earlier) version of the national test to be equated; creating - in theory at least - matched groups, whose outcomes could be compared directly. Thus in the comparisons in each subject at each key stage, each child took only one version of the test, minimising the testing burden.

Two alternative experimental designs had been rejected, and it may help to explain why. The first would have addressed the 'natural' question, 'would children taking different versions of a test get the same results on each?' quite literally - by asking the same group of children to attempt both versions of the test. We might for instance have asked children in England to take a past version just before or after their own live tests. But (apart from the problem of finding children whose teachers had not employed the past version involved to help prepare their pupils) this would inevitably have created unequal conditions between versions. Children would be more highly motivated when taking their operational version than when taking another for research purposes, so negating fair comparisons. This same problem makes it difficult for national test development agencies to make comparisons with operational results for equating purposes when trialling future versions of national tests. We shrank from the obvious variant of asking children to take two different non-operational versions of a national test as well as their own live version. This would have narrowed the time gaps we could address and finding schools willing to submit their pupils to so great a testing burden in the name of research might well have proved difficult, especially if they had to refrain from using 'last year's' test to prepare their pupils.

The second alternative would have been to use unmatched groups, taking both one or other of the versions being compared and some other concurrent measure of achievement in the subject concerned. The latter would estimate the differences in ability between the experimental groups - in short, an anchor test strategy. Here the difficulties would have included both finding a suitable concurrent measure and extending the testing burden. To be wholly satisfactory an anchor test probably needs to be as long as the test being compared, which would incur the testing burden above for less advantage. Substantial variations in the distribution of ability between the groups might also have challenged many of the assumptions behind the statistical models involved in equating, making the results potentially insecure, even given a suitable anchor measure.

But the Project did take the precaution, wherever possible, of a belt and braces policy: supplementing the equivalent groups design by collecting 'available data' which might serve the purpose of an anchor test, so that any small differences in ability between the experimental groups which might arise in practice could be monitored and allowed for. This sought the benefits of the anchor test approach without incurring its costs, in all senses of the word. As

a result we enjoyed the luxury of a supplement to the research question 'do the (equivalent) groups of children taking different versions of the test get the same results?', namely, 'not accounted for by any variations in ability we can detect between them'.

The supplementary data available

To be of use to us, any supplementary data available had to provide a recent and relevant estimate of pupils' achievement or ability relating to the subject concerned, which was common across all the schools involved. Fortunately, for many of the comparisons we were making, schools had just conducted their own end of key stage assessments, under the auspices of the Northern Ireland Council for the Curriculum, Examinations and Assessment (NICCEA), QCA's counterpart in Northern Ireland. At KS3 these included national tests for English, Mathematics and Science which were ideal for our purpose. And whilst national tests were not set at KS2, NICCEA did provide a 'catalogue' of assessment units for schools to use to support their teacher assessments in English and Mathematics, which were then moderated. These teacher assessments provided adequate estimates of ability to monitor our KS2 comparisons in these subjects, and the NICCEA KS2 Mathematics assessments were also used to compare the equivalence of the groups involved in our comparisons at KS2 in Science.

The dividing point between KS1 and KS2 comes a year later in NI than in England, so end of KS1 assessments were not yet available for the children involved in our comparisons between KS1 tests from different years. In the absence of any assessments common to all schools we instead collected details of children's date of birth as a surrogate, as there is extensive research evidence (for instance Sharp et al., 1994) that this is strongly associated with achievement at this age.

Random assignment of children to the different versions of the tests

Virtually all children[2] in the appropriate cohort within each school involved were asked to participate. The schools were asked to form experimental groups to take the different versions of the test by using spiral quasi-random assignment (Petersen et al, 1989) within gender. Simply put, this means allocating alternate boys on the school/class register to versions x and y in turn, and then repeating the process for girls. This should avoid school, teaching group and neighbourhood effects, which have been shown to affect achievement on national tests (Daniels and Stainton, 1994). Spiral assignment likewise avoided problems which might arise from streaming, selection, single sex provision etc. by distributing them equally between test forms. But, as explained above, the project did also gather data relating to gender and achievement, to provide a basis for checking the equivalence of groups assigned to different versions, and thus a means of controlling statistically for any differences which might emerge.

[2] Although schools had, and exercised, the option of excluding any for whom the materials were unsuitable.

Sample selection

The groups of children involved in each comparison needed to contain a balance of schools of varying sizes, from urban and rural locations etc., and to span the full ability range. But there was no need for them to be a strictly representative sample of Northern Ireland's children.

Sample selection and the initial approaches to schools, asking them to participate, were undertaken by NICCEA, acting on the Project's behalf. We would like to record our thanks for all NICCEA's help and assistance, without which the project would not have been possible - especially in Phase 1, when the schedule was extremely tight.

In Phase 1 a random sample of 91[3] first/ primary schools were approached, with a letter from NICCEA (drafted jointly with UCLES) describing the project and asking them to participate in the comparisons of the KS1 Reading Comprehension tests and / or either the KS2 English tests or the KS2 Mathematics tests. The response rate was a little disappointing, perhaps reflecting the short notice being offered to schools and the timing of the study, which was close to the end of term (earlier in Northern Ireland than in England) and clashed with pre-arranged school visits for Primary 7 children in many schools. But of the 91 schools approached initially, 34 agreed to participate. NICCEA then approached further 'replacement' schools by telephone, 9 of which agreed to take part (i.e. 43 in all). Of these, 41 schools agreed to administer the KS1 test, together with one or other of the KS2 tests. The remaining two schools only wished to administer a KS2 test. Further details of the samples for each of these comparisons are reported later, but the project reached its target of identifying 1000+ children to participate in each experimental comparison in this phase.

In Phases 2 and 3 separate samples of schools for each comparison being made were again initially approached by NICCEA, by letter, followed by subsequent approaches to replacement schools (normally by letter, as schedules were less tight than Phase 1) when required. Again the project's target was to identify 1000+ children to take part in each comparison. Further details of these samples are also given later, in the sections reporting each experimental comparison.

In total 11,762 children from 184 schools took part in the experimental testing programme. These included 3,304 KS1 children from 62 schools, 4,390 KS2 children from 82 schools, and 4,068 KS3 children from 40 schools.

We would like to record our special thanks to all the teachers and children who helped the Project, both for their efforts on our behalf and for the kindness and good humour we encountered in our contacts with them.

[3] An initial random sample of 150 schools was reduced to 91 by the exclusion of small schools, i.e. those with fewer than 15 pupils in their Primary 7 cohort. This practice was followed throughout, for logistical reasons.

Test administration

After the initial contact had been established by NICCEA all subsequent contacts with schools were made by the project staff.

Test materials were dispatched from UCLES offices and schools were provided with special instruction manuals for administering the tests, designed so that they could use both years' versions at the same time in the same room, in order to minimise disruption. These were produced by the project team, who borrowed directly from the operational QCA test administration manuals wherever possible.

Materials for teachers also included (and emphasised the importance of following) the Project's instructions for randomising the selection of the groups of children taking the 1996 and 1999/ 2000/ 2001 versions of the tests, using spiral allocation within gender: i.e. assigning alternate boys and girls on class registers to each version.

Teachers administered tests during the specified testing windows of 1 to 2 weeks, just after the dates of operational testing in England in each phase of the Project. Materials could not be dispatched before operational test dates for obvious security reasons. Testing took place under normal testing conditions, supervised by children's own teachers, as would live tests. Arrangements were made for carriers to collect the completed test materials from schools and deliver them to UCLES for marking.

Marking and data preparation

Marking teams of suitably experienced teachers were established. To ensure that marking followed the proper procedures, these were briefed by senior members of the operational QCA marking teams, with experience of supervising external markers in both years being compared wherever possible. The operational marking schemes used in both years being compared were followed rigorously throughout.

Scripts from each school were randomly apportioned amongst the markers, to ensure that inter-marker variation could not be confounded with test form: each marker having equivalent samples drawn from the sets of scripts from each school for every component in both versions of the test. Statistical monitoring of the patterns of marks awarded by each marker was undertaken and we can be confident that the marking process was conducted fairly and accurately in each case, without bias with respect to the versions of the test being compared.

Data entry included marks on each question; together with supplementary information (name, teaching group, gender, date of birth (for KS1 only) and National Assessment Level achieved - for KS2 and KS3) from class registers (for KS1) or the school's KS2 or KS3 National Assessment Record Forms. Total marks for each component etc. were then computed, so that arithmetic

errors by markers were avoided. Data checks included investigation of cases where the numbers of children in a given school and teaching group assigned to the two versions of the tests varied by more than one, to check that allocation to test forms appeared random. Means and standard deviations of marks within schools were also checked during this process. Where random assignment appeared to be in doubt, or whenever schools had informed us that they had used 'alternative' approaches, data from the class(es) concerned were excluded from the analyses comparing test forms which follow; as failure to randomise would have compromised the research design.

Some children were found to have completed one component from one version and another from the other year being compared, due to errors in test administration. Data for all such cases were excluded, as were data for their classmates if there were several such cases in a teaching group (as in these circumstances the remainder could not be assumed to form random groups). All children who were 'partially absent' were also of necessity excluded from comparisons, but data from their classmates were included, on the assumption that absence was not systematically related to assignment to test forms.

Details of the final samples of children for whom valid data were obtained are included with the reports on each comparison which follow.

2.2 KS1 Reading Comprehension experimental comparison: 1996 v 1999

Historical trends

Table 2.2.1 shows the percentages of children awarded each level in KS1 Reading national assessments (including sub-divisions a to c of level 2) throughout the period 1997 to 1999. In 1997 74% of children reached level 2 and by 1999 this percentage had grown, albeit fairly modestly, to 79%.

Table 2.2.1 KS1 Reading Comprehension 1996-1999 (% at each level*)

Year	*W*	*Level 1*	*Level 2C*	*Level 2B*	*Level 2A*	*Level 3*
*1996***	-	-	-	-	-	-
1997	-	-	17%	17%	14%	26%
1998	3%	-	16%	16%	19%	26%
1999	-	-	16%	18%	16%	29%

* Rows do not total 100% as absentees and children who were disapplied are not included.
** Not recorded, as the test was not a statutory requirement in 1996.

No equivalent figures are available for 1996 as the KS1 Level 2 Reading Comprehension test was not a statutory requirement then - it remained optional until 1997. The experimental comparisons we can make are necessarily limited to comparing the 1996 and 1999 versions of this KS1 Reading test: i.e. to the reading test standards alone. Thus, because tests were not the only elements determining national assessments in this period, this cannot address the wider question 'have KS1 Reading national assessment standards shifted'?

The validity of experimental comparisons in schools in Northern Ireland

Variations in style and content between the 1996 and 1999 tests
The 1996 and 1999 KS1 Level 2 Reading Comprehension tests were closely scrutinised by project staff, who considered the types of stimuli and questions and the skills tested by each question.

These two versions of this test included a similar range of types of question and stimuli. Apart from one question in 1999, all were single mark questions. A narrative stimulus appeared in each test. In 1996 the text was a simple familiar story theme set in a realistic context; whereas the 1999 narrative text was more complex, imaginative and set in a less familiar context. An information / reference text was included in both years, as well as an information text (an invitation) in 1996, and an instructional text (a recipe) in 1999.

The focus of the 1996 items was mostly literal comprehension and information retrieval, with only one item appearing to require inference and deduction. Many 1999 items also addressed information retrieval and literal comprehension. However, Part 1 of the 1999 test included 3 items requiring an explanation and 3 items requiring inference, one of which required the expression of opinion.

Two 1999 questions were based on a labelled diagram, whereas this format did not feature in 1996.

The overall impact of the differences between the tests, particularly in relation to the skills required, seemed likely to make the 1999 test more difficult than the 1996 test. But such variations in absolute difficulty are the reason why cut-scores are adjusted from year to year and would not inhibit our experimental comparisons. Given that our concern is that experimental comparisons should be unbiased, the key question is, 'are there features of one version or other of the test which would be especially easy or challenging for children from Northern Ireland'? To answer this we have also to consider how the NI and English school systems differ.

Structural and curricular issues

Structural issues

In England KS1 spans ages 4-7 years – the reception class and years 1 and 2 (Y0, Y1, Y2): whilst in Northern Ireland it extends a further year to include ages 4-8 years – termed Primary 1 to 4 (P1, P2, P3, P4). In both systems KS2 concludes at age 11, spanning Y3, Y4, Y5 and Y6 in England and P5, P6 and P7 in Northern Ireland. Thus the NI children in P3 who took the KS1 tests had not reached the end of 'their' KS1, although they had enjoyed a similar length of schooling to English children of the same age. Reaching the end of key stage 1 landmark might have some effect on achievement levels if teaching and learning is adapted to it. But there seems no reason why this should invalidate our comparisons, as children taking different versions of the KS1 Reading Comprehension test would be equally affected.

In England teacher assessments of levels achieved in each attainment target were reported alongside test results. In Northern Ireland statutory assessments at the end of KS1 were by teacher assessment only, supported by the use of a selection from a catalogue of assessment units distributed by NICCEA, who also moderated schools' assessment portfolios. Both systems used similar systems of (ten in all) national curriculum 'levels', with similar expectations of progression. Levels are awarded on a 'best-fit' basis to summarise achievement. Target setting and benchmarked comparisons of schools' achievements in national tests were introduced in England in 1998, but were not a feature of primary education in NI. So whilst the two systems were in many ways similarly structured, it appears likely that the children involved in the KS1 experimental comparisons may have been somewhat less exposed to formal testing than children of the same age in English schools.

But even if some P3 children in NI are less test-wise than many Y2 children in England, the design of the study should ensure that matched groups are assigned to the different test forms, so that the comparisons between versions remain valid. Thus again there seems no reason why the differences in assessment arrangements observed should bias our comparisons between test forms from different years.

Curricular issues

Project staff also reviewed the official KS1 curriculum documentation, to compare the primary curriculum in English and Mathematics in England and Northern Ireland. In both England and NI the curriculum is defined within programmes of study and attainment targets. Expected levels of performance are set out in level descriptions for each attainment target. The programmes of study at KS1 cover up to level 3 and it is expected that the majority of children will be working at level 2 at the end of the KS in both systems. The programmes of study at KS2 cover up to level 5 and it is expected that the majority in NI will be working at either level 3 or level 4 at the end of KS2. The majority of English children are expected to be working at level 4 at this stage.

The curriculum documents from England and Northern Ireland relating to reading comprehension at KS1 are not easily compared, varying greatly in layout and presentation. England's documentation tends to be more detailed and specific.

For instance it lists categories of reading materials (e.g. 'stories and poems from a range of cultures…') and features the reading material should contain (e.g. 'illustrations that are visually stimulating and enhance the words of the text'). But we have little reason to expect that in practice children in the two systems experience different ranges of reading materials.

Similarly the NI curriculum contains an overarching statement stipulating the need for children to be able to use picture, phonic, contextual and other cues in order to identify unfamiliar words. Three other brief statements also refer to children understanding the structure of texts, noticing the construction of words and spellings and being able to develop a sight vocabulary. England's curriculum is again far more detailed and includes a breakdown of the knowledge, skills and understanding to be acquired in terms of phonic, graphic and grammatical cues, word recognition strategies and contextual understanding. But such variations in style cannot be taken to mean that the detailed content specified in England is not valued and taught by teachers in Northern Ireland. It may be tempting to speculate that teaching and learning experiences in NI may be more variable than in England, but much depends on the extent to which England's teachers use curriculum documents to determine their daily lesson planning.

Level descriptors from the two curricula cover similar ground in the main. There are two noteworthy exceptions. In England children at level 2 are expected to be able to offer opinions about themes and events in a range of texts (a requirement reinforced in the National Literacy Framework for the term in which the KS1 tests are administered). No such 'personal response' is required in Northern Ireland at this stage, although by level 3, children are expected to 'recognise' some of the main points by selecting information from the text and to use inference and deduction to appreciate its meaning. In this latter respect NI's requirements outdo England's, where such skills do not appear until the next key stage. Achievement on some questions in the 1996 and 1999 versions of the KS1 Reading Comprehension tests might have been affected if such curriculum details do govern what children learn, but it is not

easy to guess the balance of advantage for the groups assigned to the two versions. In 1999 question 12 from the narrative text asked why the two friends were quiet for a long time. Would children from NI have been more likely to be fazed by this? Or would those from England be more likely to be fazed by the level of inference demanded? Questions 3 & 5 from 1999's narrative text (against only question 14 from the narrative text in 1996) also seem to require inference. Would we expect children from NI to perform especially well in 1999 - bearing in mind that teachers might only have alerted their most able children to such matters?

The National Literacy Framework details what is to be covered at word, sentence and text level during each term in England. Under reading comprehension (non-fiction) in Year 2, term 1, it indicates that children should be taught to 'read simple written instructions in the classroom, simple recipes, plans, instructions for constructing something; to note key structural features, e.g. clear statement of purpose at start, sequential steps set out in a list, direct language.'. Northern Ireland's curriculum does not explicitly refer to instructional text and if children there have little experience of it they may have found the 1999 test (which contained such material) more difficult than their counterparts in England.

However this might be counterbalanced by the inclusion of more questions requiring inference/deduction in 1999, for which children from NI might perhaps be better prepared.

Teachers' views
Fieldwork in the schools taking part in the study (using semi-stuctured interview schedules which asked teachers about the match between the tests and their teaching) during visits to deliver feedback to schools, suggested that teachers thought the 1996 and 1999 Reading Comprehension tests were non-threatening. This perhaps goes some way towards balancing the point that the Primary 3 children tested were not used to formal testing. Teachers appreciated the inclusion of fiction and non-fiction. In general the tests were felt to reflect the curriculum and reading schemes used in Northern Ireland.

Some teachers pointed to context issues. The 1996 test used the name Raj (fiction questions 1-18) which might be unfamiliar to many children and introduced trains, which are not found in the south-west of the province. But such issues might also arise in many rural schools in England.

Teachers considered the 1999 test likely to prove harder than the 1996 test because of the inclusion of questions requiring inference and deduction, rather than literal comprehension (as had project staff). Some also suggested that their children might be unfamiliar with reading information from a diagram and with instructional texts.

In summary
How may we summarise this evidence relating to the potential effects of curricular and structural issues on performance on the 1996 and 1999 versions of the KS1 Level 2 Reading Comprehension test? We have noted

that some differences between the two systems do not challenge the validity of our comparisons because the experimental design assigns matched groups to take the two versions of the test. It does not matter if the NI children involved perform better or worse than children from England might have done. We are not seeking to compare achievement in the two systems. Our interest lies only in comparing the two versions of the KS1 Reading test. We are primarily interested in curricular factors etc. which might bias this comparison.

Two differences between the 1996 and 1999 tests may be of some importance. The 1999 test contains a set of four questions based on an instructional text which children from Northern Ireland might not have been prepared for (and in the event the NI children involved here did not perform well on these questions). However it also included a higher proportion of questions demanding inference, which teachers in Northern Ireland might have been more likely to address, and in practice children in the Northern Ireland schools involved did perform relatively well on at least some of these inferential questions.

These potential effects within the 1999 test would seem to counterbalance one another and so it seems reasonable to use a Northern Ireland sample to compare KS1 Level 2 Reading Comprehension test standards.

The data

A random sample of 91 schools (excluding schools with cohorts of less than 15 pupils) were approached on the Project's behalf by NICCEA and asked to participate. The initial approach was made within the first week of the Project's lifespan, but was still closer to the date of testing than would have been ideal, so that many schools had committed themselves to school visits in the testing window proposed. Nonetheless their response was outstanding, with 32 of the schools agreeing to take part, together with an additional 9 'replacement' schools contacted later by telephone, making 41 schools in all with a total of 2,243[1] Primary 3 children.

Test materials and full instructions for test administration (versions of QCA's instructions, edited by project staff to facilitate simultaneous administration of the two versions) were supplied to the schools, including arrangements for the spiral allocation of random groups of children to the 1996 and 1999 versions of the test. Testing took place in late May 1999, soon after operational testing was finished in England. In the event all schools bar one (a special school which withdrew after seeing the materials and judging them too demanding) returned completed tests to UCLES in Cambridge. Schools were also asked to supply details of each P3 child's date of birth[2], teaching group and gender.

[1] In this first phase of the Project, primary schools were asked to provide P3 children to take KS1 L2 Reading Comprehension and also P7 children to take either KS2 English or KS2 Mathematics; hence the relatively large KS1 Reading Comprehension sample.

[2] KS1 does not end until P4 in NI, so end of key stage assessments were not available. DoB provides an alternative means of checking the equivalence of experimental groups.

The KS1 Reading Comprehension tests were marked by a team of 4 suitably experienced KS1 teachers, briefed by a KS1 teacher with experience of assessing KS1 throughout the period 1996 -1999. Scripts from each school were randomly apportioned amongst the markers, to ensure that inter-marker variation could not be confounded with test form. The operational marking schemes for 1996 and 1999 were followed rigorously throughout and marking took place at UCLES Research & Evaluation Division's offices, where the marking team was supervised by project staff with relevant teaching experience. Statistical monitoring of the marks awarded revealed that variations in the marks awarded by different markers were insignificant. We can thus be confident that the marking process was conducted fairly and accurately with respect to the comparison of the two versions of the test.

Data entry included marks for each question, together with name, gender, teaching group and date of birth. Total marks were computed, checked against markers' totals and differences were reconciled. Further data checks included the numbers allocated to each version and mean scores within class groups, to establish that these were consistent with random allocation to test forms. In a small number of cases this appeared to be in doubt and data from these teaching groups were excluded from subsequent analyses.

All the analyses reported below are based on the 1,860 children for whom valid reading comprehension test scores were available, together with details of their gender and date of birth. This discounts the data from 40 children whose age details (and in three cases gender details too) were missing.

Were the groups assigned to the 1996 and 1999 versions equivalent?

Table 2.2.2 compares the distributions of children's ages, in months, of the pupils with valid experimental test results in the groups taking the 1996 and 1999 versions of the KS1 Reading Comprehension tests. The differences were not statistically significant, suggesting that the groups were reasonably well matched in this respect, which we would expect to correlate with achievement for children of this age (Sharp et al, 1994).

Table 2.2.2 Age (in months) by test version

	Age															
	<84	*84*	*85*	*86*	*87*	*88*	*89*	*90*	*91*	*92*	*93*	*94*	*95*	*>95*		
1996	1.3	9.5	7.8	8.9	8.5	6.0	8.9	8.1	9.4	8.2	10.0	6.8	5.9	0.6	100%	n 949
1999	0.9	8.0	8.3	7.2	9.8	5.6	8.5	7.8	7.4	8.3	9.3	9.9	7.8	1.1	100%	n 911

n 1,860 *Chi-square Likelihood Ratio 29.21, df 13, significance 0.11*

Table 2.2.3 shows the numbers of boys and girls in the groups taking each version. The gender balance of the group taking the 1996 version was more even than that of the group assigned to the 1999 version, where boys slightly outnumbered girls, although the differences were not statistically significant.

Table 2.2.3 Gender by test version

	Girls	*Boys*	*Total*	
1996	49.6	50.4	100%	n 949
1999	47.0	53.0	100%	n 911

n 1,860 *Chi-square Likelihood Ratio 1.31, df 1, significance 0.25*

How do results on the 1996 and 1999 test forms compare?

Table 2.2.4 shows the overall distributions of Reading Comprehension test levels achieved by the groups assigned to the 1996 and 1999 versions.

The distributions achieved suggest that the group taking the 1999 version of the test tended to achieve lower levels than those assigned to take the 1996 version, a finding which was statistically significant.

Table 2.2.4 Levels achieved, by test version

	NCL <2	*NCL 2c*	*NCL 2b*	*NCL 2a*		
1996	13.5	20.5	18.8	47.2	100%	n 949
1999	16.6	22.8	23.8	36.8	100%	n 911
n 1,860	*Chi-square Likelihood Ratio 21.76, df 3, significance <0.0001*					

Given that girls frequently achieve higher scores than boys, on average (Gipps & Murphy, 1994; Johnson, 1996), it was possible that the differences in gender composition of the groups assigned to the two versions might be influencing the comparison between years. However, table 2.2.5 shows the distributions of levels achieved by gender as well as test version and confirm that both boys and girls assigned to the 1999 version tended to achieve lower levels than those of the same gender taking the 1996 test.

Table 2.2.5 Levels achieved by gender and test version

		NCL<2	*NCL 2c*	*NCL 2b*	*NCL 2a*		
(a) Girls	1996	9.3	17.4	20.8	52.4	100%	n 471
	1999	12.1	19.4	26.2	42.3	100%	n 428
n 899	*Chi-square Likelihood Ratio 9.77, df 3, significance 0.02*						
(b) Boys	1996	17.6	23.6	16.7	42.1	100%	n 478
	1999	20.5	25.9	21.7	31.9	100%	n 483
n 961	*Chi-square Likelihood Ratio 11.44, df 3, significance 0.01*						

If we take the psychometric liberty of assuming that the National Curriculum 'level scale' (below 2, 2c, 2b and 2a) is an equal interval scale scored 0,1,2,3 or 4, we can represent the above information in terms of mean 'sub-levels', and use analysis of variance (ANOVA) to examine the differences in achievement between genders and groups assigned to the two versions. Table 2.2.6 displays the means and the ANOVA results. These confirm that girls achieved significantly better (and less widely spread) results than boys, as has often been reported elsewhere and that those assigned to the 1999 version achieved significantly lower (sub-)levels than those taking the 1996 form. A further analysis of covariance (ANCOVA), controlling for the small variations in the distributions of age between the 1996 and 1999 groups, yielded the same conclusions.

It thus seems fair to conclude that the levels achieved by children assigned to the 1999 test were significantly lower than those achieved by those assigned to the 1996 test. As there seems no valid reason to deny that these experimental data were a fair test of the null hypothesis, it would appear that

the standards set by the 1999 level 2 Reading Comprehension Test differed from those in the 1996 version.

Table 2.2.6 Mean transformed KS1 RC NCLs by gender and test version

		Mean	*SD*	*n*
1996		2.00	1.10	949
	boys	1.83	1.16	478
	girls	2.16	1.02	471
1999		1.81	1.11	911
	boys	1.65	1.13	483
	girls	1.99	1.05	428
Total		1.90	1.11	1,860

ANOVA (n 1,860)
Gender F = 43.17, df 1, significance <0.001
Test Version F = 12.59, df 1, significance <0.001
Interaction of Gender & Version F = 0.002, df 1, significance 0.96

ANCOVA *(n 1,860)*
Age F = 7.01, df 1, significance 0.015
Gender F = 44.44, df 1, significance <0.001
Test Version F = 13.38, df 1, significance <0.001
Interaction of Gender & Version F = 0.01, df 1, significance 0.93

How might standards in 1996 and 1999 differ?

The means and standard deviations of the marks achieved by boys and girls taking the two versions of the test are shown in table 2.2.7. In both versions girls outscored boys in total Reading Comprehension marks (by 2.36 marks, on average, on the 1996 test and by 2.85 on the 1999 version). For both sexes mean marks on the 1996 test were significantly higher than on the 1999 test – by 1.19 marks for girls and 1.68 for boys – despite the fact that the 1999 test had a maximum mark of 31 compared to the 28 marks available in the 1996 version. We can conclude that in absolute terms the 1999 test proved more difficult than the 1996 version. What then was the effect of the cut-scores set?

Table 2.2.7 KS1 Reading Comprehension (Level 2): Means and Standard Deviations

	Girls			*Boys*		
	n	*mean*	*sd*	*n*	*mean*	*sd*
1996 max 28	471	21.91	7.05	478	19.55	8.18
1999 max 31	428	20.72	7.95	483	17.87	8.76

ANOVA (n 1,860)
Gender F = 48.74, df 1, significance <0.001
Test Version F = 14.86, df 1, significance <0.001
Interaction of Gender & Version F = 0.43, df 1, not significant

Table 2.2.8 shows the cut scores for the award of levels 2a – 2c in both 1996 and 1999. The 1999 cut-scores for levels 2b and 2c (but not 2a) were slightly lower than their 1996 equivalents, suggesting that those responsible for setting standards had detected the variation in difficulty our experiment revealed. But had the 'correct' allowance been made?

Table 2.2.8 also includes the outcomes of test equating. Linear equating (where equivalent marks correspond to the same number of standard deviations above or below the mean - Angoff, 1971) was not appropriate in this instance as the distributions for the two test versions were of very different shapes, even though both were positively skewed. In 1996 the distribution was truncated, with the maximum mark also the mode. Instead column three shows the 1999 marks deemed equivalent to the 1996 thresholds in equipercentile equatings for boys and girls (separately, in view of the gender imbalance and distributional differences between the two groups). This non-linear approach defines scores as equal if they correspond to the same percentile rank in the equating group. It differs from linear equating if the test distributions are not the same shape but may suffer from floor and ceiling effects because, by definition, the distributions converge towards the minimum / maximum scores.

Table 2.2.8 KS1 Reading Comprehension (level 2) cut-scores and equated marks

KS1 RC L2	*1996 Cut-Score (max 28)*	*1999 Cut-Score (max 31)*	*1999 mark equated to 1996 Cut-Score (Equipercentile)*
Level 2a	25	25	Girls 24 / Boys 24
Level 2b	21	19	Girls 17 / Boys 17
Level 2c	10	9	Girls 8 / Boys 10

In these data, the equipercentile equatings of the two versions' distributions suggested that the 1999 cut-score for levels 2a might have been set a mark too high and that for 2b two marks too high for equivalence with the 1996 version. The equipercentile equatings for 2c in boys and girls yielded different results. For boys the 1999 2c cut score appears about one mark too low, whilst for girls it seems about one mark too high; so it seems reasonable to conclude that it was about right.

Regrettably, these experimental data do not in themselves provide any basis on which to assess the effect size of such differences on national test results.

We should also re-state the point that the 1996 test was optional, like that set in 1995. Because of this we must repeat the point that these experimental comparisons are confined to the equivalence of the standards of the two tests in question. They could not address the equivalence of the operational standards applied in national assessments of KS1 Reading in 1996 and 1999, which involved other assessments.

2.3 KS1 Mathematics experimental comparison: 1996 v 2000

Historical trends

When looking at achievement at KS1 in Mathematics it is essential to recognise the structure of the assessment 'package' involved. This includes two elements. One is a Mathematics 'Task', which is designed to be suitable for the assessment of children working towards or achieving at Level 1. The second is the 'Test', with which these experimental comparisons are concerned. Only children who might hope to achieve Levels 2 or 3 are required to take the tests, which are targeted at this higher range of attainment. Thus KS1 Mathematics national assessment results are determined by the combined effects of both the Task and Test set each year. Even though Levels 2 and 3 can only be gained via success on the Tests, fluctuations in results between years might be caused by variations in the numbers of children asked to take each year's test; a decision taken by their teachers. Given some reservations about its interpretation arising from the above, table 2.3.1 shows the trend in KS1 Mathematics results over time.

Table 2.3.1 KS1 Mathematics results 1996 - 2000*

	W	*Level 1*	*Level 2c*	*Level 2b*	*Level 2a*	*Level 3*
1996	3%	15%	19.5%	19.5%	24%	19%
1997	2%	14%	18%	20%	25%	20%
1998	2%	12%	23%	24%	18%	19%
1999	3%	10%	23%	22%	20%	21%
2000	2%	7%	17%	23%	25%	25%

* rows do not total 100% as absentees and children disapplied are not shown

Clearly these do not portray uniform improvement throughout the ability range. Whilst the proportion of the cohort reaching Level 2 has risen steadily year after year, the proportion reaching Level 2a (or better) declined between 1997 and 1999, but jumped to an all-time high in 2000.

The validity of experimental comparisons in schools in Northern Ireland

Variations in style and content between the 1996 and 2000 tests
Project staff reviewed the 1996 and 2000 KS1 Mathematics tests in detail, considering the types of questions and stimuli and the topics and skills required by each question.

They were broadly similar, although

- The 2000 test included one question where children must complete a table and another where they should show their working out - neither type being included in the 1996 version.
- The first five questions on the 2000 version were administered orally by teachers: no such provision was included in 1996.
- There were only three data handling questions in 2000, compared to six in 1996.
- In the 2000 version, more items used multiplication and division operations, and fewer used addition and subtraction.

- Thirteen questions relating to number manipulation or properties were included in the 2000 version, compared to only ten in 1996.
- Fractions only featured in the 2000 version, whilst assessment of the position and movement of shape only featured in 1996.

Such differences in emphasis only affect the validity of our comparisons between test versions if these variations in style and content also relate to curricular differences between Northern Ireland and England. Performance by children from NI on one or other version of the test might then be affected when achievement by children from England would not have been affected in the same way. This issue is addressed below.

Curricular and structural issues

Structural issues relevant to KS1 have already been discussed thoroughly in 2.2 above. The arguments there are as relevant to mathematics as to Reading Comprehension and need not be repeated in full here. In brief, although pupils will not have reached the end of their key stage and may have less experience of formal assessment, this in itself should not bias comparisons between versions of QCA tests and does not rule out experimental comparisons involving P3 children in Northern Ireland.

Detailed review of official documentation revealed that the KS1 Mathematics curriculum is on the whole quite similar in England and NI. However there are some significant curricular differences that should be considered. The NI curriculum has five attainment targets in Mathematics at KS1 compared to England's three. NI's ATs 3 & 4 (Measures and Shape & Space) correspond to England's AT3, whilst NI's AT5 (Data handling) is not a separate attainment target at KS1 in England. But in both curricula ATs relate directly to sections within their respective programmes of study, which cover much the same content. And some apparent disparities (e.g. use of computer software, including a database) are resolved by their inclusion in the introduction to the NI programme of study, rather than the programme itself. In general the Level Descriptors in NI and England also correspond, although some items do appear at different levels. In a few instances emphasis does vary sufficiently to suggest that they might affect children's test performance. Those topics which seem potentially problematic and which feature in the 1996 or 2000 tests are detailed below.

Number

- England's curriculum details 'division' needing to be understood as repeated subtraction.
- Using halves and quarters is classed in Level 2 in England and Level 3 in NI.

Shape, space and measures

- The term 'faces' is used in NI to replace 'surfaces' used in England.
- England's curriculum specifies that children should recognise hexagons and pentagons and know their geometric features, whilst NI's does not include such details.

National Numeracy Framework
The NNF for teaching mathematics was introduced in England in September 1999. This highlights key objectives which teachers should prioritise, including the following items not detailed in the NI curriculum:

- Know and use halving as the inverse of doubling.
- Use a ruler to draw and measure lines to the nearest centimetre.
- Choose and use appropriate operations and efficient calculation strategies to solve problems, explaining how the problem was solved.

But of course NI teachers may well still teach such things. The items listed above will only impact on performance to the extent that teaching is affected by curriculum documentation.

Overall, this curricular review gave no reason for concern regarding the variations in question style and type described initially. There seems no reason why children from Northern Ireland should find any of the questions particularly inaccessible. Close analyses of the two versions of the test suggested that although there were a few questions on each where children from England or NI might have an advantage, the balance of relative advantage seemed fairly even, and no different in one version of the test than the other. The latter is the key issue, as our concern lies with the comparison between versions, not the level of achievement of children in Northern Ireland or any comparison with performance levels in England.

The principal concern might be that NI's longer KS1 (by a year) leads to some differences in the sequencing of teaching, so that in some schools the teaching of some topics may have been postponed beyond the end of P3, when children took part in this study. This might affect children's readiness to address questions on any such topics. The potential effects of this cannot be predicted from desk analyses as they stem from individual teachers' classroom practice. Teachers' views on these two versions of the KS1 mathematics tests are therefore highly relevant.

Teachers' opinions
In Phases 2 and 3 of the Project, which included KS1 mathematics, teachers in each school were asked to complete questionnaires showing which test questions, if any, they thought their children might find difficult or problematic, given their educational experience, and why. Nineteen of the twenty-one schools returned completed questionnaires. In a few cases their responses were clarified in the course of telephone fieldwork, during follow-up work with the schools.

Table 2.3.2 shows which questions elicited comments (and the number of schools so commenting) suggesting that the material they tested might not have been fully covered by a school's teacher(s). The table also shows the questions which teachers suggested they would expect their children to find difficult. Whilst the latter is in principle a very different issue, which could apply equally in England, it is possible that some respondents might have used 'difficult' as short-hand for poor preparation. The level of concern in these

regards was considerable, as table 2.3.2 indicates. Difficulties arose primarily because, in P3, some NI schools tend to work only with numbers less than 100 and/or delay multiplication and division until P4.

Table 2.3.2 Questions 'not covered' or 'expected to prove difficult' by teachers

1996 Questions	*n schools saying not covered*	*n schools saying difficult*		*2000 Questions*	*n schools saying not covered*	*n schools saying difficult*
1				*1*		
2				*2*		
3				*3*		
4				*4*		
5				*5*		
6				*6*		
7				*7*	3	1
8				*8*		
9				*9*		
10	10	1		*10*		
11				*11*		
12				*12*		
13				*13*	7	5
14	6	1		*14*		
15	11	1		*15*		4
16	7	4		*16*	7	1
17				*17*		2
18	4	1		*18*	13	1
19	3	1		*19*	3	2
20	11	3		*20*	8	3
21	12	3		*21*	6	1
22	10	4		*22*	11	2
23				*23*	8	2
24	6	6		*24*	7	1
25	9	3		*25*	2	3
26	8	2		*26*	8	3
27	5	5		*27*	7	4
28	5	5		*28*	11	3
29	12	3		*29*	6	2
30	8	5		*30*	9	2
				31	11	1
				32	7	3
				33	5	2
				34	5	2
Total n Comments	127	48			144	50
Qns affected	16/30	16/30			20/34	22/34

This central curricular issue affected a substantial number of questions in both the 1996 and 2000 versions of KS1 Mathematics, but fortunately the evidence from teachers' responses to the questionnaire suggested that the impact of this issue on the two versions of the test seemed to be fairly even. A similar proportion of the questions from the 1996 and 2000 versions were cited as being affected, in respect of both coverage and difficulty. If this is so, we might expect P3 children from NI to achieve lower marks on these tests than children of the same age and ability from England[1]. But their performance should still provide a guide to the relative difficulty of the two versions of the test, as the 1996 and 2000 versions of the KS1 Mathematics tests both appear to be affected by this problem to a similar extent.

[1] It is difficult to be confident about such matters. To the extent that time is released by not covering some topics, NI children might be expected to be better on others, which would be emphasised by their teachers instead, to fill the time available for learning.

The data

A random selection of primary/first schools[2] in Northern Ireland, including schools of different sizes and locations, was approached by NICCEA on the Project's behalf and invited to participate. Twenty-one schools agreed and administered the 1996 and 2000 versions of the KS1 Mathematics test to their Primary 3 children in June 2000, shortly after the operational test dates in England. In each school spiral allocation within gender groups was used to form random groups of boys and girls taking the two versions of the test. Because P3 children are still a year away from the completion of Northern Ireland's key stage 1, school assessments on a 'common' basis across schools were unavailable and the only ancillary data collected was the pupils' date of birth - which was provided for 1016 of the 1061 children who completed the tests. Teachers administered the tests to their own pupils, using a version of QCA's operational instructions modified by the project team to facilitate the administration of two versions of the test.

The scripts were returned to Cambridge where they were marked by a team of three KS1 teachers with experience of operational marking, briefed and co-ordinated by a team leader with KS1 assessment experience spanning the period 1996-2000. Marking was carried out in UCLES' offices and markers were each allocated random selections of the scripts from each school. Statistical quality control checks indicated that there were no significant differences between the marks awarded by the three markers. Subsequent data entry and cleaning involved range and total checks and investigation of the numbers of boys and girls from each school, together with the means and standard deviations of test scores within schools, suggested that the schools had all followed (the inevitably somewhat complex) spiral allocation instructions correctly.

Were the groups assigned to the two versions equivalent?

Table 2.3.3 shows how the proportions of boys and girls in the groups assigned to the 1996 and 2000 versions of the test were fairly even (a Chi-Square test proving not statistically significant), although a slightly higher proportion of boys took the 2000 version than the 1996 version. Given the likelihood of gender differences in performance this suggests that, if only as a precaution, gender should be taken into account in subsequent data analyses.

Table 2.3.3 Gender by test version

	Girls	*Boys*	*Total*	
1996	49.6	50.4	100%	n 546
2000	48.7	51.3	100%	n 515
n 1,061	*Chi-square Likelihood Ratio 0.77, df 1, n.s.*			

The distributions of age (in months at June 2000) for the two groups were similar (see table 2.3.4). Of children for whom date of birth was available, those assigned to the 1996 version had a mean age of 7 years 5.6 months

[2] Schools with less than 15 children in the cohort were not approached, for logistical reasons.

whilst those assigned to the 2000 version had a mean age of 7 years 5.5 months. The groups appear well matched in this respect.

2.3.4 Age (months) by test version

	Age														
	<84	*84*	*85*	*86*	*87*	*88*	*89*	*90*	*91*	*92*	*93*	*94*	*95*	*>95*	
1996	4.3	6.3	7.2	6.1	8.8	8.6	8.2	8.6	6.9	7.6	8.8	7.8	6.5	4.8	100% n 546
2000	3.2	7.7	10.2	4.5	9.4	6.7	8.1	9.8	5.5	6.1	8.4	8.6	9.0	2.8	100% n 515

n 1,016 (missing data = 45) *Chi-square Likelihood Ratio 33.72, df 13, n.s*

How do results on the 1996 and 2000 test forms compare?

Table 2.3.5 shows the percentage of children achieving each level on the two versions of the test and table 2.3.6 reveals the results for boys and girls separately.

Table 2.3.5 Levels achieved by test version

	NCL <2	*NCL 2c*	*NCL 2b*	*NCL 2a*	*NCL 3*	
1996	11.5	30.0	24.0	25.3	9.2	100% n 546
2000	11.1	34.4	31.8	18.8	3.9	100% n 515

n 1,061 *Chi-square Likelihood Ratio 24.06, df 4, significance <0.001*

Although similar numbers of children reach level 2 or higher in the two versions, better results are obtained by the group allocated to the 1996 version, on which more children obtained level 3 or 2a.

Table 2.3.6 Levels achieved by gender and test version

		NCL<2	*NCL 2c*	*NCL 2b*	*NCL 2a*	*NCL 3*	
(a) Girls	1996	9.2	31.4	24.0	25.5	10.0	100% n 271
	2000	8.4	33.5	32.7	21.5	4.0	100% n 251

n 522 *Chi-square Likelihood Ratio 11.5, df 4, significance <0.05*

(b) Boys	1996	13.8	28.7	24.0	25.1	8.4	100% n 275
	2000	13.6	35.2	31.1	16.3	3.8	100% n 264

n 539 *Chi-square Likelihood Ratio 14.01, df 4, significance <0.01*

This pattern held for both boys and girls.

Assuming equal interval properties for the NCL scale, we transformed the levels achieved to numeric form (<2 = 1.67, 2c = 2.0, 2b = 2.33, 2a = 2.67, 3 = 3.0), enabling the calculation and comparison of means reported in table 2.3.7. Analysis of variance (ANOVA) suggested that the differences in the distributions of levels achieved by the groups allocated to the two versions were significantly different, as were the gender differences. Analysis of covariance (ANCOVA), further controlling for any variations in the ages of boys and girls and/or pupils assigned to the two versions, confirmed these findings.

If we accept that these analyses are a fair test of the null hypothesis (which does not seem unreasonable) we must therefore conclude that the Levels achieved via the 1996 version of the KS1 Mathematics test were higher than those achieved by the group taking the 2000 version.

Table 2.3.7 Mean transformed KS1 Mathematics test NCLs by gender and test version

		Mean	*SD*	*n*
1996		2.30	0.39	546
	boys	2.29	0.40	275
	girls	2.32	0.39	271
2000		2.23	0.34	515
	boys	2.20	0.34	264
	girls	2.26	0.33	251
Total		2.27	0.37	1,061

ANOVA	*Gender F = 4.28, df 1, significance <0.05*
(n 1,061)	*Test Version F = 9.05, df 1, significance <0.01*
	Interaction of Gender & Version F = 0.34, df 1, n.s.
ANCOVA	*Age F = 12.36, df 1, significance <0.001*
(n 1,061)	*Gender F = 4.17, df 1, significance <0.05*
	Test Version F = 7.74, df 1, significance <0.01
	Interaction of Gender & Version F = 0.28, df 1, n.s.

How might standards in 1996 and 2000 differ?

Overall, mean levels reported in table 2.3.7 were 0.07 higher for the group assigned to the 1996 test. But we need to know more about the ways in which the tests and their associated cut-scores compare.

In terms of absolute difficulty, as represented by raw marks, the 1996 test was in fact the more difficult of the two - the mean mark obtained being 13.3 compared to 14.4 on the 2000 version. The spread of marks was fairly similar - standard deviations being 5.7 and 5.5 respectively. But it is the combined effects of test difficulty and the cut-scores set which determine access to Levels. It would seem that the 2000 level setting decisions were correct in perceiving that the test was easier but over-compensated for this (at least in contrast with 1996 - a version of the test those responsible would never have considered in practice) by setting cut-scores which were even higher than necessary.

What changes to the 2000 cut-scores would have been required to bring them into line with their equivalents from the 1996 version of the test? Table 2.3.8 presents results from an equipercentile equating of the distributions of total marks on the two versions of the test. It is worth noting that determining equivalent mark points necessarily involves some approximation, especially when comparatively short mark scales are involved, as is the case here, with relatively high proportions of children 'at' each point on the scale.

Table 2.3.8 KS1 Mathematics cut-scores and equated marks

KS1 Mathematics	*1996 Cut-Score*	*2000 Cut-Score*	*2000 mark equated to 1996 Cut-Score (Equipercentile)*
Level 3	22	25	22
Level 2a	16	19	17
Level 2b	12	14	13
Level 2c	7	8	7/8

It would appear that the misalignment of the cut-scores was greatest at the Level 3 threshold, where the 'best' equivalent would have been 22 marks, the same as was taken in 1996, three marks less than the threshold of 25 actually set. At Level 2a the 2000 threshold would have needed reducing by two marks to be 'equivalent' to the 1996 version, whilst a reduction of only 1 mark would be required to have brought the threshold at levels 2b into line with 1996 standards. Arguably, the 2c threshold might have been allowed to stand, or perhaps have been reduced by a single mark.

Although the proportion reaching the 'expected level' might not be affected, such changes would be likely to generate substantial improvements in the finer details of national results, compared with those produced by the operational thresholds for 2000. Levels 2a and 3 would be most affected. Unfortunately these experimental data do not provide any basis for estimating the size of such effects.

So, it would appear that in KS1 Mathematics tests, test standards relating to Level 3 and the sub-divisions within Level 2 might have been raised between the 1996 test and the 2000 version. The rise in the proportions of children awarded Levels 2a and 3 nationally in 2000, would seem more than justified. Indeed the thresholds set for these abler children might have been somewhat severe. In addition, the Level 2c thresholds in 1996 and 2000 represent a consistent view of standards, in effect validating the higher proportions of children recorded as reaching this (the expected) Level in the more recent year. The evidence of this study suggests that schools have indeed been successful in improving the quality of children's work in KS1 Mathematics between 1996 and 2000.

These experimental data do not help us to determine when the disparities at the higher thresholds might have been introduced. Level setting decisions in any or all of 1997, 1998, 1999 or 2000 might have contributed. But the historical pattern of results shows that access to the higher range of levels was restricted in 1998 and it seems possible that the additional demands on abler children may stem from then.

2.4 KS2 English experimental comparisons: 1996 v 1999 & 1996 v 2000

Historical Trends

The percentages of children in England achieving each level in the KS2 English tests set from 1996 to 2000 are shown in table 2.4.1. This records a remarkable improvement in national results, with the percentage achieving level 4 or better rising from 57% in 1996 to 75% in 2000 and the proportion reaching level 5 more than doubling over the same period.

Table 2.4.1 Key Stage 2 English 1996-2000 (% at each level*)

Year	*Below 3*	*Level 3*	*Level 4*	*Level 5*
1996	9%	30%	45%	12%
1997	7%	26%	47%	16%
1998	6%	26%	48%	17%
1999	7%	20%	48%	22%
2000	6%	17%	46%	29%

* rows do not total 100% as absentees and children disapplied are not shown

2.4.1 Initial experimental comparison: 1996 v 1999

The validity of experimental comparisons between the 1996 and 1999 versions of the KS2 English tests in schools in Northern Ireland

Variations in style and content between the 1996 and 1999 versions

Reading
A similar range of question types was used in the two versions of the 45 minute (+ 15 minutes reading time) reading test. However, there were more questions which required longer answers in the 1996 test than the 1999 version, although the 1999 test included more multiple mark items.

In the 1996 version, stimuli included an information / reference text, an instructional text with a diagram and a narrative text. In 1999 the stimuli included an information / reference text with pictures and captions, a poem with reference information, a glossary extract, an introduction and a cartoon. More subjectively, the project team considered the presentation of the 1999 stimuli materials more lively (and, equally arguably, the themes themselves more likely to appeal to children) than the 1996 version.

There were 2 parts to the 1996 version and 4 parts to 1999. Part 4 of the 1999 test included items drawing on the whole booklet, addressing text type, theme and purpose, and which required textual reference. In 1996 a higher proportion of questions required explanation, opinion and inference, whereas in 1999 there was greater emphasis on a wider range of skills and knowledge, including vocabulary; imagery; language; purpose; structure and presentation; text types and themes.

The project team's view was that the greater diversity in skills assessed in 1999 might make this version the more difficult, although the effects of

diversity on difficulty may be offset against the greater emphasis on opinion, explanation, inference and deduction in 1996.

Writing
Both versions of the Writing test lasted 45 minutes (+ 15 minutes reading). The 1996 version gave children a choice of one writing task from a range, including 3 narratives and 1 newspaper article. In 1999 the range available included 2 narratives, 1 persuasive letter and 1 persuasive leaflet. A common planning sheet was provided for all narratives in 1996, with a separate planning sheet for the newspaper article. In 1999 there were separate planning sheets for each alternative stimulus.

The purpose of the writing was emphasised in the more extensive 1999 stimulus materials, which included more effective (in our view) picture prompts than the 1996 version. Emphasising the purpose of writing for the letter and leaflet in 1999 could have helped pupils to target writing more successfully.

Writing prompts for 1999 were more extensive, which may have helped some pupils by providing more support. Conversely, the extra reading might have disadvantaged pupils with reading problems, although there is no evidence of this and there were 'preventative measures' in place: teachers read the booklet with pupils at the beginning of the test and 15 minutes reading time was provided before writing began.

Although the criteria applied in the mark scheme remained the same, there were changes to the way marks were awarded between 1996 and 1999. The mark schemes provide criteria against which to award marks for 'Purpose and Organisation' (maximum 21); 'Punctuation' (maximum 7); and 'Style' (maximum 7). For each of these, descriptive criteria were provided for six quality levels in 1996. By 1999 these had been reduced to five by the removal of the lowest category. Markers used the criteria by deciding which best applied and awarding the appropriate mark. Note here that for all three elements in Writing, more marks were available than there were categories, and any intervening marks between the specified marks for each category were not used (see table 2.4.1.1 below).

The effect of removing the lowest category for 1999 and adjusting the marks awarded was that there were differences in the marks awarded for work which would have fallen into the two lowest categories for 1996, as shown below:

Table 2.4.1.1 Writing: marks awarded for given criteria in 1996 and 1999

	Purpose & Organisation		*Punctuation*		*Style*	
	1996	*1999*	*1996*	*1999*	*1996*	*1999*
High Level 5	21	21	7	7	7	7
Level 5	18	18	6	6	6	6
Level 4	15	15	5	5	5	5
Level 3	12	12	4	4	4	4
Below Level 3	6	9	2	2	2	2
Well below Level 3	3	0	1	0	1	0

The impact was that for Purpose and Organisation three more marks were awarded to answers falling 'just below level 3' in 1999, although any falling

'well below' would get no credit, rather than the three marks they would have been awarded in 1996. For Punctuation and for Style the effect was simply that any falling 'well below level 3' in 1999 would get no marks instead of the one awarded in 1996. The different mark schemes for the two years were applied rigorously by markers in this experiment, but these changes would have little effect on our analyses, as very few children score in this range.

Spelling and Handwriting
Spelling: Both versions of this test required 10 + 5 minutes. The Spelling tests for 1996 and 1999 were administered in the same way, with the teacher reading a passage and pupils spelling words in that context.

A comparison, based on initial consonant blends and vowel phonemes, indicated that the words in the two tests were similar with respect to these. However the 1999 list included more words with a large number of syllables.

Handwriting: This skill was assessed in the same way in both years.

Structural and curricular differences between England and Northern Ireland

Project staff reviewed official documents describing primary schools and the English curriculum in England[1] and in Northern Ireland, with a view to identifying any features which might bear unevenly on the versions of tests set in different years, and so bias our experimental findings.

Structural issues
In England KS1 spans ages 4-7 years – the reception class and years 1 and 2 (Y0 - Y2): whilst in Northern Ireland it extends a further year to include ages 4-8 years – termed Primary 1 to 4 (P1 - P4). In both systems KS2 concludes at age 11, spanning Y3 - Y6 in England and P5 - P7 in Northern Ireland. Thus the NI children in P7 who took the KS2 English tests had reached the end of 'their' KS2, having enjoyed a similar length of schooling to English children of the same age, albeit differently split between KS1 and KS2.

In England externally set and marked tests were set at the end of KS2 but in Northern Ireland statutory assessments at KS2 were by teacher assessment, supported by the use of a selection from a catalogue of assessment units distributed by NICCEA, who also moderated schools' assessment portfolios. Both systems used similar national curriculum 'levels' to report achievement. Target setting and benchmarked comparisons of schools' achievements in national tests were introduced in England in 1998, but are not so far a feature of primary education in NI.

Despite the absence of national tests at the end of KS2 in NI, children there did encounter 'similar' external assessments. Any NI children wishing to transfer to selective grammar schools took (optional) transfer tests in English,

[1] This included the potential impact of the national literacy strategy in England.

Mathematics and Science which were externally set and marked. In 1998/9 these transfer tests were taken by 67% of 11 year olds, so a majority of children in NI were likely to be familiar with formal test conditions. Even if some were less test-wise than others, the design of our study should have ensured that groups assigned to different test forms were matched in this respect, so that the comparisons between test versions remained valid. Thus again there seems no reason why the differences in assessment arrangements observed should bias our experiment.

Curricular issues
In both England and NI the curriculum is defined within programmes of study and attainment targets. Expected levels of performance were set out in level descriptions for each attainment target. The programmes of study at KS2 cover up to level 5 and it was expected that the majority in NI would be working at either level 3 or level 4 at the end of KS2. The majority of English children were expected to be working at level 4 at this stage.

Reading: The programmes of study for reading in England and NI covered much the same areas and expected outcomes. The NI curriculum was more detailed and provided illustrative examples of the range of work to be covered and the types of activities suggested. Overall, level descriptions covered the same criteria, although some NI criteria appeared at the next higher level in the English curriculum.

Some specific points from the NI curriculum were not explicit in the English curriculum, but could be seen as subsumed within more general statements or in more detailed criteria in the National Literacy Framework (piloted in England in 1997/8 and introduced from September 1998). These include
- *modelling writing on forms encountered in reading*
- *attempting to reconstruct texts*
- *beginning to talk about the perceived intention of the author.*

Conversely there were aspects of reading detailed in the English curriculum which do not feature specifically in NI documentation, including:
- *the ability to distinguish between fact and opinion*
- *the use of figurative language.*

But these too could be seen as subsumed under general headings.

Although the use of figurative language could be seen as subsumed under the more general headings in the NI curriculum, the fact that it was not explicitly mentioned could perhaps lead teachers there to emphasise poetry less than their counterparts in England. In 1999 the reading test included questions based on a poem, several of which addressed the use of figurative language. This might have made the 1999 version of the KS2 English test less accessible for children from NI as compared with those from England.

Although there were other differences between the two reading curricula, comparison of the questions and texts involved in the 1996 and 1999 tests, suggested that they would not affect children's performance differentially.

Writing: Writing covers broadly the same areas and expected outcomes in NI and England. Both curricula included the need to write for a variety of purposes.

Handwriting: In the Handwriting test credit was given for 'joined writing', which is explicitly required in the English curriculum. This criterion was not specified in the NI curriculum. But this difference will have the same impact on performance in the 1996 and 1999 versions and thus should not invalidate comparisons between them.

National Literacy Framework: The NLF introduced in England sets out teaching targets for each term. Although NI children cover the same targets there was no such schedule. If a test includes questions relating to topics scheduled to be taught in England around the time children are to be tested; this scheduling might result in a form of bias. These questions might be expected to be easier for English children, for whom such learning should be fresh in mind. This assumes that teachers do adhere to the teaching schedule set out in the framework. The Framework indicates that the work to be covered in reading comprehension during term 2 of Y6 includes the study of poetry and there should be an emphasis on constructive argument, expression and appealing to audience in non-fiction. During the same term, work to be covered for writing composition includes the construction of effective, persuasive arguments for a particular audience and harnessing the views, interests and feelings of the audience. In Term 3 of Year 6 reading comprehension includes an emphasis on poetry, including style, theme, format, language. Writing composition includes a comparison of texts in terms of style, strengths, weaknesses, values and appeal to a reader, as well as writing poems linked by theme or form.

Thus, the emphasis on poetry study during the months prior to the 1999 test in England should have led to more focused teaching on an area which featured heavily in the 1999 KS2 reading test - where the poetry section carried 18 marks. Likewise the Framework's schedule for writing could also have made the 1999 writing test easier, in relative terms, for English pupils, since in the run up to the test period it emphasised persuasive writing, which featured more strongly in the writing prompts in 1999 than 1996. NI children may not have enjoyed these curricular emphases, so handicapping their performance on the 1999 version, relative to their English counterparts.

Teachers' opinions

Curriculum documents might be misleading if teachers' practice does not reflect such apparent variations. What did teachers in the schools involved think? After testing, fieldwork by project staff in schools included a series of questions asking if pupils' learning experiences might have made them unfamiliar with some questions and/or especially well prepared for others; as well as discussing more general issues.

Several teachers remarked that children often required more time than was available to complete the tests, especially in Reading, where some felt that the quantity of material was excessive. In England children would be likely to

be better prepared for the tests. But there was no suggestion that one version might have been more affected than the other by this issue.

A number of comments suggested that children might not be familiar with questions which required them to offer a personal response or an explanation or opinion, which was especially apposite for the 1996 test. The formal assessment of handwriting was also unfamiliar.

The 1999 spelling test was thought likely to prove far more difficult than the 1996 version, but there was no suggestion that one or other version was more or less fair for pupils from Northern Ireland.

In summary
This review suggests that children in NI may be less familiar with poetry and figurative language (included only in the 1999 reading test) and thus perhaps making this version more difficult (relative to the 1996 test) for our experimental sample than it would have been for children in England in 1999. More arguably, the NLF's schedule might confer a similar advantage in relation to persuasive writing, again making the 1999 test (which contained two such options in the choice of four offered, as compared with one of the four in 1996) harder for children from NI than those from England. Both these potential effects might depress performance (in our experiment) on the 1999 version, making it look relatively severe in our comparisons. This might be considered when looking at the outcomes of the experiment, but teachers made no reference to these issues and did not suggest their pupils might find such questions difficult. Classroom practice may not vary to the extent the documentation might imply.

A few teachers did suggest that their NI pupils might have been unfamiliar with questions asking for a personal response or an explanation, which were more common in the 1996 version. This would tend to counterbalance the potential effects described above.

Overall there seemed to be many more similarities than differences and the experimental comparisons planned seemed feasible.

The data

A random sample of schools (excluding schools with cohorts below 15, for logistical reasons) were approached (on the project's behalf) by NICCEA and asked to participate. In total 21 of the schools asked to administer the KS2 English test (with a total of 1,094 P7 children on roll) agreed to do so.

Test materials and full instructions for test administration (versions of QCA's instructions, edited by project staff to facilitate simultaneous administration of the two versions) were supplied to the schools, including arrangements for the spiral allocation of random groups of children to the 1996 and 1999 versions of the test. Testing took place in late May 1999, soon after operational testing was finished in England. All these schools returned completed tests to UCLES

in Cambridge, together with details of each child's gender and their end of KS2 assessments in English; the latter conducted according to NICCEA's instructions and subject to their monitoring arrangements.

The KS2 English tests were marked by a team of 3 suitably experienced KS2 teachers - briefed by a senior member of the operational KS2 English marking team and co-ordinated by a team leader, both with experience of marking KS2 English throughout the period 1996 -1999. Scripts (for both the 1996 and 1999 versions) from each school were randomly apportioned amongst the markers, to ensure that inter-marker variation could not be confounded with test form. The operational marking schemes for 1996 and 1999 were followed rigorously throughout and marking took place at UCLES Research & Evaluation Division's offices, supervised by project staff with primary experience. Statistical monitoring of the marks awarded revealed that variations in the marks awarded by different markers were insignificant, confirming that marking was conducted fairly and accurately for both versions of the test.

Following data entry, total marks were computed and checked against markers' totals and differences were reconciled. Further data checks included the numbers allocated to each version and mean scores within schools, to establish that these were consistent with random allocation to test forms. In a small number of cases this appeared to be in doubt and data from these teaching groups were excluded from subsequent analyses.

All the analyses reported below are based on the 844 children for whom complete sets of data were available; providing their performance on the 1996 or 1999 KS2 version of the KS2 English tests, their gender and their Northern Ireland KS2 Assessments for English.

Were the groups assigned to the 1996 and 1999 test forms of equivalent ability?

Table 2.4.1.2 shows the NI KS2 English achievements by the groups assigned to the 1996 and 1999 versions of QCA's KS2 English test. The groups were remarkably similar on this criterion, showing that random assignment to test forms had worked well in this instance; creating well matched groups. Having acknowledged this, we can note that the (statistically insignificant) differences present suggest that if anything the group taking the 1999 version was very slightly inferior to that taking the 1996 test.

Table 2.4.1.2 NI KS2 English Assessment Levels by test version

	NI NCL 2	*NI NCL 3*	*NI NCL 4*	*NI NCL 5*		
1996	3.3	20.5	56.4	19.8	100%	n 420
1999	3.5	22.4	55.2	18.9	100%	n 424
n 844	*Chi-square Likelihood Ratio 0.54, df 3, n.s.*					

How do results on the 1996 and 1999 test forms compare?

Let us now ask the key question, how did the groups' KS2 English levels match? Table 2.4.1.3 reveals that they clearly do not match closely, as the null hypothesis would expect. Instead the results obtained by the 1999 group were markedly better than those obtained by the 1996 group, with 9.8% more children reaching level 4 and 10.3% more reaching level 5 than in the 1996 group. These differences are highly significant and indicate that, on the basis of these comparisons, standards in these two versions might differ.

Table 2.4.1.3 KS2 English Levels by test version

	NCL <2	*NCL 2*	*NCL 3*	*NCL 4*	*NCL 5*		
1996	3.1	1.4	30.2	56.2	9.0	100%	n 420
1999	1.7	1.7	21.7	55.7	19.3	100%	n 424
n 844	*Chi-square Likelihood Ratio 24.02, df 4, significance <0.001*						

If we again take the statistical liberty of assuming that these national curriculum levels (below 2, and 2-5) may be represented as 0 and 2-5 respectively, on a scale assumed to enjoy equal interval properties, we can use mean levels to summarise these data and at the same time consider a further factor, namely gender.

Table 2.4.1.4 presents the breakdown of means by year and gender and includes results from an analysis of variance (ANOVA) which confirmed that the gender differences and the differences observed between test versions were statistically significant. An analysis of covariance (ANCOVA), using Northern Ireland KS2 English assessments (where again girls tend to outperform boys) as a control variable, also confirmed the importance of the variations between test versions.

Table 2.4.1.4 KS2 English mean transformed NCLs by gender and test version

		Mean	*SD*	*n*
1996		3.64	0.90	420
	boys	3.45	0.96	209
	girls	3.82	0.81	211
1999		3.88	0.85	424
	boys	3.73	0.91	230
	girls	4.05	0.74	194
Total		3.76	0.89	844

ANOVA	*Gender F = 31.92, df 1, significance <0.001*
(n 844)	*Test Version F = 18.48, df 1, significance <0.001*
	Interaction of Gender & Version F = 0.17, df 1, n.s.
ANCOVA	*NI English NCL F = 500.57, df 1, significance <0.001*
(n 844)	*Gender F = 13.73, df 1, significance <0.001*
	Test Version F = 33.58, df 1, significance <0.001
	Interaction of Gender & Version F = 0.00, df 1, n.s.

Girls' mean levels were about a third of a level higher than boys', and whilst the average levels achieved on the 1999 test version were almost a quarter of a level higher than those for children taking the 1996 version, we should

recognise that the smaller proportion of girls taking the 1999 test may affect this estimate of the difference in standards between the two versions.

Again, if we can assume that there is no valid reason to deny that these experimental data are a fair test of the null hypothesis, these comparisons would lead us to conclude that the standards set in the 1996 and 1999 versions of the KS2 English tests were different. Levels achieved by children assigned to the 1999 test were significantly better than those achieved by those assigned to the 1996 test, other factors being equal. Whilst our review of the curriculum and tests did leave us with some reservations about the validity of the null hypothesis in KS2 English, on balance these suggested that children taking the 1996 test might be advantaged by comparison with those assigned to the 1999 version. Given that results suggest the opposite, we have no reason to doubt the validity of the differences identified empirically.

How did children perform in the components in each year's test?

Table 2.4.1.5 presents the means and standard deviations of marks on each component in the 1996 and 1999 tests, including a breakdown by gender.

Table 2.4.1.5 Component means & standard deviations of marks in 1996 and 1999

		All	*n = 420*	*Girls*	*n = 211*	*Boys*	*n = 209*
		mean	*sd*	*mean*	*sd*	*mean*	*sd*
1996	*Reading*	30.69	9.55	32.82	8.86	28.54	9.76
	Spelling	7.74	2.44	8.02	2.38	7.45	2.47
	Handwriting	3.43	0.90	3.55	0.93	3.30	0.86
	Writing	19.49	4.74	20.74	4.32	18.22	4.82
	Writing-Purpose	11.66	3.19	12.44	2.92	10.88	3.27
	Writing-Style	3.94	1.01	4.15	0.94	3.73	1.03
	Writing-Punctuation	3.88	1.24	4.15	1.14	3.61	1.29
	KS2 English Total	61.34	15.19	65.12	14.32	57.51	15.12
		All	*n = 424*	*Girls*	*n = 194*	*Boys*	*n = 230*
		mean	*sd*	*mean*	*sd*	*mean*	*sd*
1999	*Reading*	26.41	8.70	27.67	8.15	25.34	9.01
	Spelling	6.84	2.75	7.11	2.61	6.60	2.84
	Handwriting	3.30	0.99	3.48	1.01	3.14	0.95
	Writing	20.54	4.03	21.43	3.52	19.79	4.28
	Writing-Purpose	12.39	2.56	12.87	2.34	11.99	2.67
	Writing-Style	4.16	0.98	4.30	0.88	4.04	1.04
	Writing-Punctuation	3.99	1.16	4.26	0.96	3.76	1.26
	KS2 English Total	57.08	14.27	59.69	13.19	54.88	14.79

ANCOVA (Handwriting)	*Covariate (NI NCL) F = 115.78, df 1, significance <0.001*
(n 844)	*Gender F = 9.88, df 1, significance <0.01*
	1996 v 1999 F = 2.90, df 1, n.s.
	*Interaction Gender * Year F = 0.94, df 1, n.s.*
ANCOVA (Writing)	*Covariate (NI NCL) F = 294.29, df 1, significance <0.001*
(n 844)	*Gender F = 30.59, df 1, significance <0.001*
	1996 v 1999 F = 22.25, df 1, significance <0.001
	*Interaction Gender * Year F = 1.76, df 1, n.s.*

The gender differences in these data are readily apparent and similar to those recorded elsewhere (Gipps & Murphy, 1994; Johnson, 1996). Girls obtained

higher mean marks than boys on every component (and the sub-elements of Writing) on both versions of the test. On every component (except Handwriting) the boys also exhibited a wider spread of marks than girls.

Marks for Reading were, on average, about 4 marks higher for the 1996 version than 1999, and the same pattern holds for Spelling, where mean marks in 1999 were about 1 mark lower. These differences are not unexpected; being largely[2] the product of each version's test development process. Such variations in scores between tests for different years are the reason why different cut-scores are set each year, as we should expect the cut-scores to compensate for differences in test difficulty. We cannot meaningfully compare the raw marks from different versions directly, without taking the effects of cut-scores into account. Whilst levels for Reading and Writing were reported from 1997 onwards at KS2, these were not provided in 1996. Consequently separate cut-scores for Reading and Writing are not available for 1996, precluding meaningful direct comparisons or equating of scores on the Reading test.

The case for direct comparison of marks for Handwriting and Writing is arguable. The criteria used to assess both were the same in each version, although we have already explained that there were some small adjustments to the marks awarded for Writing to lower categories of work between 1996 and 1999. Handwriting marks awarded to the group taking the 1999 version of the test were, on average, a little lower (by about a tenth of a mark) than those obtained by the group taking the 1996 version. But the average Writing mark was higher for the 1999 group than the 1996 group (by about one mark), with differences within all three elements of Writing (Purpose, Punctuation and Style) conforming to the same pattern.

In an analysis of covariance, which took the effects of gender and ability (estimated by Northern Ireland National Assessments) into account, differences between the Writing marks obtained on the two test versions proved statistically significant (see Table 2.4.1.5). In a similar analysis, differences between 1996 and 1999 means for Handwriting just failed to reach statistical significance.

So, on the assumption that adjustments to the mark schemes had little practical impact on the distributions of marks, it seems that children with equivalent results on NI National Assessments tended to obtain slightly better Writing marks on the 1999 version than the 1996 version of the test. The most likely explanation for this would seem to be the more extensive stimulus materials in 1999. These provided more scaffolding for the tasks and emphasised the purpose of the writing (which features strongly in the Northern Ireland curriculum), as well as providing better graphics and separate planning sheets, each specifically designed for a given writing task.

Fortuitously perhaps, this difference in Writing marks counter-balances the slightly lower mean marks for spelling obtained by the 1999 test group.

---

2 but potentially confounded with any variations in spelling ability between the groups taking the two test forms

The effects of question choice in Writing

In each version children had a choice of four options for their writing task. In 1996 these consisted of three narrative writing tasks (stories) and an informative writing task (an article). In 1999 there were two narrative tasks and two informative writing tasks (a letter and a leaflet). Table 2.4.1.6 shows the numbers of boys and girls in our experiment choosing each title from each version, together with the means and standard deviations of marks achieved. Analyses of covariance which investigated the statistical significance of the effects of question choice and gender within each year (whilst controlling for the effects of ability via Northern Ireland National Assessments) are also reported.

Table 2.4.1.6 Writing options

	n	*mean*	*sd*	*female (mean / n)*	*Male (mean / n)*
1996					
1 Look who's talking (narrative/story)	294	19.44	4.71	20.56 / 148	18.31 / 146
2 No time to lose (narrative/story)	50	19.78	4.69	21.52 / 25	18.04 / 25
3 The longest day (narrative/story)	46	18.76	5.45	20.95 / 21	16.92 / 25
4 School days (informative/article)	30	20.57	3.83	20.88 / 17	20.15 / 13
Overall	420	19.49	4.74	20.74 / 211	18.22 / 209
1999					
1 Spider supporter (informative/letter)	80	20.86	4.37	22.55 / 31	19.80 / 49
2 Sea world (informative/leaflet)	124	20.62	3.68	21.68 / 50	19.91 / 74
3 If pictures could speak (narrative/story)	128	20.86	3.69	21.34 / 71	20.26 / 57
4 Home at last (narrative/story)	92	19.70	4.54	20.45 / 42	19.06 / 50
Overall	424	20.54	4.03	21.43 / 194	19.79 / 230

ANCOVA 1996 (n 420)	*Covariate (NI NCL) F = 156.06, df 1, significance <0.001* *Gender F = 7.63, df 1, significance <0.01* *Question Choice F = 0.64, df 3, n.s.* *Interaction of Gender & Q. Choice F = 0.76, df 3, n.s.*
ANCOVA 1999 (n 424)	*Covariate (NI NCL) F = 174.48, df 1, significance <0.001* *Gender F = 14.64, df 1, significance <0.001* *Question Choice F = 1.29, df 3, n.s.* *Interaction of Gender & Q. Choice F = 0.48, df 3, n.s.*

In the groups taking both the 1996 and 1999 versions of the papers, girls score higher marks than boys, irrespective of the question chosen. Gender effects are statistically significant in both versions.

The patterns of question choice are far from even. In the 1996 version the first option (a narrative) was by far the most popular, attracting 70% of all children, whilst the fourth option (the only non-narrative) was selected by only 7%. In 1999 the first option (a non-narrative question) attracted 19% of children, the second (also non-narrative) 29%, the third (narrative) 30% and the fourth (narrative) 22%. The two non-narrative questions in 1999 were thus far more popular than the single non-narrative available in 1996. These 1999 non-narratives were of course the first two in order of presentation and they proved especially attractive to boys.

There appear to be some variations in the mean scores achieved by candidates selecting different options. But in both versions these differences are not in fact large enough to be statistically significant (after allowing for gender effects and variations in the ability of the groups selecting each option) and could easily have arisen by chance. It would seem likely that, in this experiment at least, the markers proved up to the challenge of scoring each of the options available to the same standard. There is thus no evidence here to suggest that optional questions might have been marked to different standards or have otherwise varied in difficulty. Question choice does not seem to have contributed to any differences in test standards between the 1996 and 1999 forms.

How might standards in 1996 and 1999 differ?

Table 2.4.1.7 shows the cut-scores providing the thresholds for each National Curriculum Level in both the 1996 and 1999 versions of the test. This reveals that compared to 1996, 1999 thresholds were two marks lower at NCLs 2 and 3, and nine marks lower at NCLs 4 and 5. The standard setting process seems to have discerned that the 1999 version was, on balance, more difficult than that set four years earlier – concurring with our empirical comparison of scores. We should however remember that the standard setting process never in fact compared these two tests directly as we have done, as it instead concentrates on a series of year on year decisions aiming to carry forward the previous year's standards. Thus standard setting in 1997 tried to carry forward standards from 1996; likewise 1998 from 1997 and in turn 1999 from 1998. What can we say about the outcome of this series of incremental decisions, given the wisdom of hindsight and the information afforded by our direct experimental comparison? Earlier analyses have suggested that standards in 1996 and 1999 were not equivalent. By how much might they differ?

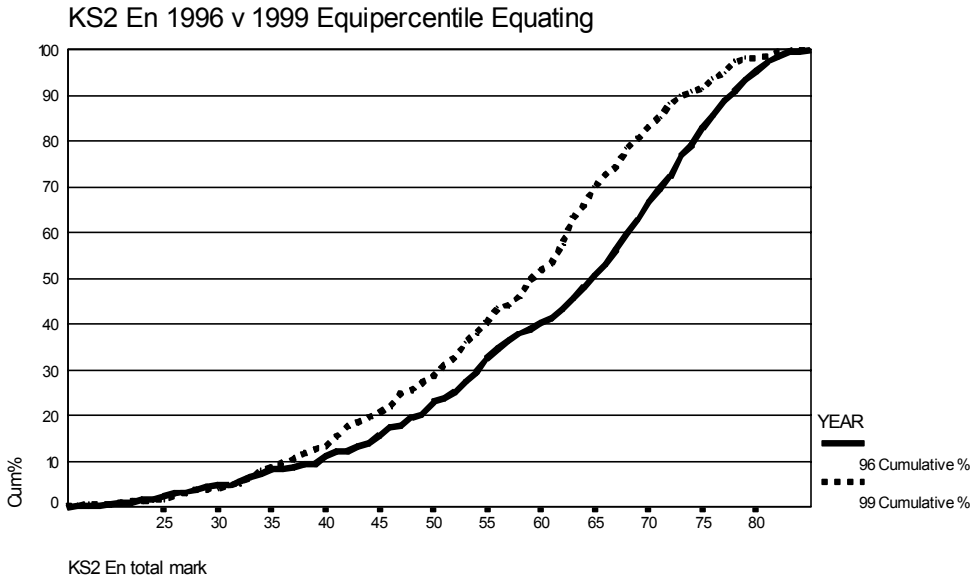
Table 2.4.1.7 also includes the outcomes from two alternative methods of test equating. Column three shows the 1999 marks considered equivalent to the 1996 cut-scores in a linear equating (i.e. the mark corresponding to the same number of standard deviations above or below the mean - Angoff, 1971). Column four shows the 1999 marks deemed equivalent to the 1996 thresholds in an equipercentile equating[3]. This non-linear approach defines scores as equal if they correspond to the same percentile rank in the equating group. It differs from linear equating if the test distributions are not the same shape and may suffer from floor and ceiling effects because, by definition, the distributions converge towards the minimum/ maximum scores. Figure 2.4.1.1 illustrates the equipercentile equating. It displays the cumulative percentage frequency distributions of the scores obtained by the samples taking the 1996 and 1999 versions and illustrates how they necessarily converge towards the tails of the distributions.

[3] Separate equipercentile equatings for boys and girls were also carried out but have not been reported, as the very small numbers in the tails of these distributions (especially around the level 2 and 3 thresholds) led to gaps in distributions which created anomalous results. At the level 4 threshold the 1999 scores equated to the 1996 thresholds for boys and girls were fairly close (53 for girls and 55 for boys).

Table 2.4.1.7 KS2 English cut-scores and equatings

KS2 English	*1996 Cut-Score*	*1999 Cut-Score*	*1999 mark equated to 1996 Cut-Score (Linear)*	*1999 mark Equipercentile equated to 1996 Cut-Score (all children)*
NCL 2	27	25	(24.4) 24	27
NCL 3	30	28	(27.3) 27	31
NCL 4	57	48	(52.9) 53	54
NCL 5	79	70	(73.8) 74	76

Figure 2.4.1.1 KS2 English: Equipercentile Equating



Equipercentile equating suggests that to bring standards in 1999 into line with those pertaining in 1996 the 1999 cut-scores would have needed to have been set two and three marks higher at levels 2 and 3 respectively, and six marks higher at levels 4 and 5. Linear equating suggests that smaller adjustments might be required at level 5 (four marks adrift) and 4 (five marks adrift) and that 1999's Level 2 and 3 cut-scores might even be too high. However, figure 2.4.1.1 reminds us that the level 2,3 and 5 thresholds are all located at the extremes of the distributions where the linear model might be expected to be suspect and where the equipercentile equating itself is most unstable (a feature which must also make operational equating difficult). The two models are largely agreed about the extent of divergence at the key level 4 threshold: it seems that the 1999 NCL 4 cut-score would need to be about five marks higher to match 1996 standards. This does not seem unreasonable, given that the mean marks of these two well matched groups differed by about four marks, whilst the cut-scores for the two versions of the test differ by 9 marks at this level.

Where and how might standards have diverged?

Clearly we cannot say when discrepancies might have arisen. Decisions taken in the course of setting standards in 1997, 1998 and 1999 might all have contributed.

But despite being hampered by the lack of separate cut-scores for Reading and Writing in 1996, we can use a process of elimination to see where the 1996 and 1999 versions of the test diverge.

Overall, average total marks for the two (fairly well matched) groups assigned to the 1996 and 1999 versions differed by about four marks, with the higher scores being obtained by those taking the 1996 version.

But their aggregated means for Writing, Handwriting and Spelling marks (which together form the Writing element reported since 1997) are almost identical (30.66 on 1996 version v 30.68 on 1999).

So it is clear that any differences in the standards applied in 1996 and 1999 must stem largely from the combined effects of changes in the cut-scores and variations in the difficulty of the Reading component between these two versions. Aggregated Writing marks were equally difficult to gain on both versions, but whilst average Reading marks were only four marks lower on the 1999 version, the (overall) 1999 cut-scores for levels 4 and 5 were nine marks below those applied in 1996.

2.4.2 A replication: 1996 v 2000

The validity of experimental comparisons between the 1996 and 2000 versions of the KS2 English tests in schools in Northern Ireland

Variations in style and content between the 1996 and 2000 versions

Reading
Similar question types were used in the 1996 and 2000 reading tests. However the 2000 version contained only four multiple-choice questions, compared to eight in 1996, although two matching questions were included in the 2000 version, against none in 1996. In all, there were twenty-one single mark questions in 1996's test and just fifteen in 2000's.

The 1996 stimuli included an information text, an instructional text and a narrative text. In 2000 the stimuli included narrative and information texts.

There were two sections in each version. In the 1996 form of the test twelve marks were available for questions relating to the information and instruction texts (section 1) whilst the information text in the 2000 version (section 2) had a maximum total of twenty three marks. In the 1996 version, the narrative (section 2) had questions with a total of thirty eight marks, while in 2000 there were twenty seven marks available for the narrative text (section 1).

In the 2000 version there were more marks available for inference and deduction skills (9 marks) than in 1996 (6 marks). In 2000 four marks were available for purpose and authorial technique and two marks for recasting information, whereas in 1996 such matters were not included. There were three marks for organisational features in 1996, compared with only one mark in 2000. Two marks were available for vocabulary questions in 1996 compared to none in 2000 and there was one mark for a genre based question in 2000, compared with none in 1996. Overall, the main differences lay in the inclusion of questions in the 2000 version which related to purpose, authorial technique and genre, as well as a greater number of questions demanding inference and deduction.

Writing
In the 1996 version of the test children selected one task from a range of choices, including three narratives and one newspaper article. In 2000, a wider range of genres was available, including a diary, a descriptive letter (information), and two narratives. The narrative options in 2000 followed the others as questions 3 and 4; whereas in 1996 options 1, 2 and 3 were narrative. There were fewer picture prompts in the 2000 version. In the instructions for 2000, the titles were listed on the front cover with a genre cue for each one. There was also a separate planning sheet for each stimulus whereas in 1996 a planning sheet was common to all narrative prompts, with a separate planning sheet for the newspaper article. The planning sheets for the 2000 version were more detailed and closely targeted and, arguably, more supportive to the writing task.

The 2000 version's mark scheme for writing was like the one used in 1999. Although the criteria applied in the mark scheme remained the same, the

marks awarded differed slightly from practice in 1996, as described in the report on the 1996 v 1999 comparison. The net effect was likely to be small, affecting only a few less able children.

Spelling & Handwriting

The Spelling tests for 1996 and 2000 were administered in the same way, with the teacher reading the passage and pupils spelling words within that context. A comparison of the words in the tests, based on initial consonant blends and vowel phonemes, indicated that the words in the two tests were similar in this respect. However the 2000 version included more words with a relatively high number of syllables. Handwriting was assessed in the same way in both years.

The timing and structure of all components in the two versions were similar.

Curricular and structural issues

Structural variations in the organisation of schools in Northern Ireland and England have already been discussed and need not be re-visited here, as is also true of the differences between the KS2 curriculum in English in the two systems.

However inspection of the 1996 and 2000 versions of the test by project staff suggested that inherent advantages/disadvantages arising from curricular issues, as discussed earlier, might affect some questions. Specifically:

Reading
1996 reading questions which could advantage the NI pupils.
- Section 1 Q5 requires an understanding of how texts can be adapted for younger readers (3 marks).
- Section 2 Qs 4, 5, 7, 8, 11 require inference and deduction (12 marks).

2000 reading questions which could advantage the NI pupils.
- Q25 required pupils to model writing on forms encountered in reading (3 marks).
- Qs 21a, 25 require pupils to reconstruct text (6 marks).
- Q13, 14, 15, 16 require pupils to understand the perceived intention of the author (8 marks).

2000 reading questions which could disadvantage the NI pupils
- Qs 22, 25 require pupils to distinguish between fact and fiction (5 marks).
- Qs. 23, 24 require knowledge of figurative language (4 marks).

This curricular analysis suggests that the number of marks involved in questions where NI pupils could, conceivably, have an advantage, is similar in the 1996 (17 marks) and 2000 (15 marks) versions. But only the 2000 version included questions (representing 15 marks) where curricular analysis suggested the NI curriculum might create disadvantage, perhaps leading to lower marks.

Although there are other differences between the two reading curricula, comparison of the questions and texts involved in the 1996 and 2000 tests, suggests that they would not affect children's performance differentially.

National Literacy Framework effects
This too has been discussed previously in general terms. During Term 1 of Year 6 the work to be covered in reading includes the analysis of texts and writers in evoking particular responses in the reader (e.g. where suspense is well-built). NI children might thus be expected to do less well than English pupils in the 2000 version of the reading test where suspense is built throughout the story. Also in Term 1 pupils should be taught to understand how authors handle time, e.g. flashbacks, stories within stories and dreams. This could put NI children at a disadvantage, relative to English pupils, in the 1996 test - where the narrative text stimulus focuses on this area in its structure. The absence of these potential benefits from the learning experience of children in NI would seem to be counter-balanced with respect to the two versions of the test being compared and hence should not invalidate the comparisons we wish to make.

Such issues should be considered when the results of the experimental comparisons between years are evaluated. If the NI children taking the 1996 version were to obtain relatively poor results than those assigned to the 2000 version, their potential net disadvantage in reading, identified above, might provide some explanation. Much would depend upon the extent to which these differences in documentation affect teachers' planning and classroom practice.

Teachers' opinions
Questionnaires were provided for each school and sixteen (out of 20) were returned. Answers to the first question revealed that test administration had gone well. The remaining questions focussed on the fit between the test materials and the NI curriculum. Table 2.4.2.1 shows the responses to questions asking how well the versions of the KS2 English test 'reflect the content of the NI curriculum'.

Table 2.4.2.1 Fit of test versions to NI curriculum

	1996 version	*2000 version*
very closely	5	5
similar	10	10
not at all	0	0

Clearly the teachers thought both the QCA tests a reasonable match to their pupils' experience. Answers to subsequent questions confirmed this. Teachers were asked to identify any questions which would be particularly problematic to children from Northern Ireland. None of the comments made pointed to curricular issues. Few questions were deemed problematic and most comments were of a general nature, simply identifying relatively difficult questions, which would be equally true for children in England. However two teachers noted that Q23 in the 2000 reading test may have caused some difficulty - because moles are not present in Ireland!

In general the teachers liked the tests. The following comment being one of several such general plaudits: 'Tests were well structured with clear instructions. They related closely to Northern Ireland curriculum expectations'.

The data

A random sample primary schools in Northern Ireland[4] was approached by NICCEA on the Project's behalf and 21 schools agreed to participate. They administered the two versions of the KS2 English test to their P7 children in June 2000. In each school spiral assignment was used to form 'randomised groups' of boys and girls taking the two versions of the test. The 1996 version was taken by 270 girls and 309 boys, whilst the 2000 version was taken by 258 girls and 291 boys, suggesting that the groups were well balanced in this respect. Schools also supplied their own end of key stage assessments for English (using a scale similar to QCA's national curriculum levels), carried out according to NICCEA's instructions and subject to moderation.

All bar one of the schools administered the tests and their scripts were returned to Cambridge for marking by a team of three experienced KS2 markers, briefed and co-ordinated by an experienced operational KS2 marking team leader. They worked in UCLES' offices and were each assigned a random selection of scripts from both the 1996 and 2000 versions of the test. Statistical checks indicated that there were no significant differences between the marks awarded by different markers. Subsequent data entry and cleaning included range and totals checks. Investigation of the numbers of boys and girls from each school taking each version, and the means and standard deviations of scores within schools, suggested that all schools returning data had followed the instructions for spiral assignment correctly. Some children were absent for one or more test components or their Northern Ireland (NI) English assessments were not available and complete sets of data were assembled for a total of 970 children (86% of all those involved). This group provided the basis for the following analyses.

How do results on the 1996 and 2000 test forms compare?

To answer this question we must also ask the prior question 'do the groups taking the two versions provide a fair basis for comparison?'. Tables 2.4.2.2 and 2.4.2.3 show the distributions of NI English assessments for the groups assigned to the 1996 and 2000 versions QCA's KS2 English tests and the Levels achieved via QCA's tests.

The difference between the distributions of NI English assessment results achieved by the groups assigned to the different versions of QCA's test were not statistically significant. Note however that Levels 4 or 5 were reached by more of those assigned to the 1996 version (75.8%) than of those taking the 2000 version (71.6%). So if there is any suggestion of bias in the assignment

[4] Schools where the cohort was less than 15 were excluded.

process, it implies that the group taking the 2000 version were slightly inferior to those taking the 1996 version. In fact the opposite pertains in the results achieved on the QCA tests. More of those taking the 2000 version achieved Levels 4 or 5 (especially the latter).

Table 2.4.2.2 NI KS2 English Assessment Levels by QCA test version

	NI En L<2	*NI En L2*	*NI En L3*	*NI En L4*	*NI En L5*		
1996	0.0	1.8	22.4	55.8	20.0	100%	n 495
2000	0.2	1.9	26.3	50.1	21.5	100%	n 475
n 970	*Chi-square Likelihood Ratio 4.27, df 4, n.s.*						

Table 2.4.2.3 QCA KS2 English Levels by test version

	NCL <2	*NCL 2*	*NCL 3*	*NCL 4*	*NCL 5*		
1996	2.2	1.6	31.1	59.8	5.3	100%	n 495
2000	1.5	0.6	22.9	54.7	12.6	100%	n 475
n 970	*Chi-square Likelihood Ratio 52.97, df 4, significance <0.001*						

Table 2.4.2.4 shows the mean Levels obtained by boys and girls on each version of the QCA tests. Those taking the 2000 version achieved average Levels 0.28 greater than those taking the 1996 version - a similar but slightly larger difference than that evident when the 1996 version was compared with the 1999 test. Girls obtained higher average Levels than boys, as is usual (Gipps & Murphy, op.cit.), and their results also exhibited a smaller (but still very substantial) difference in means between the two test versions than the boys'. The table also includes the results of an analysis of variance (ANOVA) confirming that the gender differences and the differences between test versions were statistically significant. An analysis of covariance (ANCOVA), using NI English assessments to control for any differences between the groups assigned to the 1996 and 2000 versions, provided yet more powerful confirmation of the significance of the differences between the mean levels achieved via the two versions.

Table 2.4.2.4 QCA KS2 English mean NCLs by gender and test version

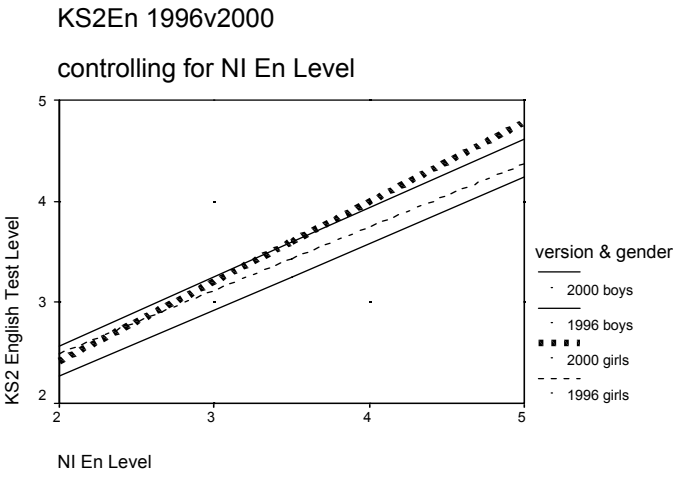
		Mean	*SD*	*n*
1996	*all*	3.62	0.80	495
	boys	3.48	0.88	259
	girls	3.77	0.67	236
2000	*all*	3.90	0.83	475
	boys	3.84	0.82	249
	girls	3.97	0.83	226
Total		3.76	0.83	970
ANOVA	*Gender F = 16.92, df 1, significance <0.001*			
(n 970)	*Test Version F = 28.63, df 1, significance <0.001*			
	Interaction of Gender & Version F = 2.10, df 1, n.s.			
ANCOVA	*NI English Assessment F = 552.29, df 1, significance <0.001*			
(n 970)	*Gender F = 7.71, df 1, significance <0.01*			
	Test Version F = 53.15, df 1, significance <0.001			
	Interaction of Gender & Version F = 1.01, df 1, n.s.			

A graphical representation of the ANCOVA is provided in figure 2.4.2.1, which shows the linear regressions of QCA KS2 English test levels on NI English assessments for the groups of boys and girls assigned to the 1996 and 2000

versions of the test. The higher levels achieved, on average, by both girls and boys assigned to the 2000 version, at any given NI English assessment level, are clear.

As was the case when the 1999 and 1996 versions were compared, the analyses of curricular and structural issues gave us no reason to expect better levels from children assigned to the later version. Indeed if anything the opposite was the case. So, given that there seems reason to assume that these data provide a fair test of the null hypothesis, we must conclude that the standards set by the cut-scores for the 1996 and 2000 versions of the tests were different.

Figure 2.4.2.1 KS2 En Levels by version and gender, controlling for NI English Assessments



How did children perform on the components in each year's test?

Table 2.4.2.5 shows the means and standard deviations of marks on each component within the 1996 and 2000 KS2 English tests, including a breakdown by gender.

Summaries of results from ANCOVA analyses, controlling for ability via NI English assessments, are presented for Handwriting and Writing. These comparisons are meaningful because the mark schemes provide continuity between years, as they do between the optional questions from which children can choose within each year. Similar comparisons for Reading and Spelling would be meaningless, because the demands of the tests set in different years can vary so much.

The girls' mean marks were significantly higher in every case, on both versions of the test. In most instances (the exceptions being Handwriting and Writing-Purpose on the 2000 version) the spread of marks was greater for boys than girls, again following the pattern usually observed.

On the 2000 version, overall, mean marks for Reading were 1.97 lower than the 1996 version's mean. Similarly, on average, Spelling marks were 1.21

lower and Handwriting marks were 0.06 lower. The latter difference was too small to be statistically significant. The higher spelling marks obtained on the 1996 version mirror the contrast with 1999 and it is perhaps notable that the project team's analyses of the spelling tests suggested that whilst the 1996, 1999 and 2000 versions were similar with respect to representation of initial consonant blends and vowel phonemes, the latter two versions contained more polysyllabic words than 1996. This may account for their proving relatively difficult.

But mean Writing marks on the 2000 version were higher than on the 1996 version, by 0.79, which was statistically significant. This difference arose largely in the marks awarded for Writing-Purpose, where the difference in means was 0.74, although marks for Writing-Punctuation were also marginally higher on the 2000 version (by 0.04 on average - not statistically significant). The difference in marks for Writing-Style was trivial.

Table 2.4.2.5 Component means and standard deviations of marks in 1996 and 2000

		All	*n = 495*	*Girls*	*n = 236*	*Boys*	*n = 259*
		mean	*sd*	*mean*	*sd*	*mean*	*sd*
1996	*Reading*	29.40	9.34	31.00	8.61	27.95	9.75
	Spelling	7.64	2.40	8.00	2.18	7.31	2.54
	Handwriting	3.49	0.79	3.62	0.75	3.37	0.82
	Writing	19.63	4.35	20.49	3.85	18.85	4.63
	Writing-Purpose	11.70	2.95	12.23	2.67	11.22	3.10
	Writing-Style	3.98	0.94	4.11	0.82	3.86	1.02
	Writing-Punctuation	3.95	1.10	4.15	0.97	3.76	1.17
	KS2 English Total	60.16	14.34	63.11	13.02	57.48	14.98
		All	*n = 475*	*Girls*	*n = 226*	*Boys*	*n = 249*
		mean	*sd*	*mean*	*sd*	*mean*	*sd*
2000	*Reading*	27.43	9.04	28.04	8.93	26.88	9.11
	Spelling	6.43	2.71	6.58	2.63	6.31	2.78
	Handwriting	3.43	0.88	3.61	0.90	3.27	0.84
	Writing	20.42	3.36	21.11	3.24	19.80	3.35
	Writing-Purpose	12.44	2.09	12.85	2.10	12.07	2.01
	Writing-Style	3.99	0.87	4.08	0.81	3.91	0.92
	Writing-Punctuation	3.99	1.05	4.18	0.92	3.82	1.14
	KS2 English Total	57.72	13.77	59.33	13.63	56.25	13.76

ANCOVA (Handwriting) (n 970)	*Covariate (NI En) F = 108.73, df 1, significance <0.001* *Gender F = 21.57, df 1, significance <0.01* *1996 v 2000 F = 0.75, df 1, n.s.* *Interaction Gender * Year F = 1.54, df 1, n.s.*
ANCOVA (Writing) (n 970)	*Covariate (NI En) F = 364.44, df 1, significance <0.001* *Gender F = 25.81, df 1, significance <0.001* *1996 v 2000 F = 17.67, df 1, significance <0.001* *Interaction Gender * Year F = 0.01, df 1, n.s.*

These variations in mean marks for Writing were similar to those observed when the 1996 version of the test was contrasted with the 1999 version. Previously, we tentatively attributed this to the provision of more extensive stimulus materials, improved planning sheets and the degree of task structuring and guidance produced in recent years. This explanation might apply equally to the contrast between 1996 and the 2000 version and will be

explored further below, when marks achieved on the various optional writing activities offered are considered.

Even the magnitude of the marks achieved replicates the previous work. The mean mark for the 1996 version in this phase (19.6) is remarkably similar to that obtained twelve months earlier when it was contrasted with the 1999 version (19.5) using a different sample of schools. The mean mark for Writing on the 1999 version was 20.5 and that for the 2000 version 20.4. This consistency enhances the credibility of these experimental contrasts.

The effects of question choice in Writing

Table 2.4.2.6 shows the mean marks obtained by those choosing each of the four optional writing tasks available in each version, together with the numbers of boys and girls selecting each title. ANCOVA analyses investigating the significance of both gender effects and differences between the marks awarded for the options available within each version, controlling for ability via NI English assessments, are also reported.

Table 2.4.2.6 Writing options

	n	*mean*	*sd*	*female (mean / n)*	*Male (mean / n)*
1996					
1 Look who's talking (narrative/story)	341	19.69	4.25	20.51/174	18.83/167
2 No time to lose (narrative/story)	88	20.09	4.23	20.44/39	19.82/49
3 The longest day (narrative/story)	40	18.23	5.12	19.64/14	17.46/26
4 School days (informative/article)	26	19.50	4.57	21.56/9	18.41/17
Overall	495	19.63	4.35	20.49/236	18.85/259
2000					
1 Moving away (imaginative/diary)	131	20.43	3.62	21.05/96	18.71/35
2 The amazing creature (descriptive/letter)	121	20.54	3.35	21.15/55	20.03/66
3 Trapped (narrative/story)	156	20.19	3.23	20.74/50	19.93/106
4 The big event (narrative/story)	67	20.72	3.18	21.96/25	19.98/42
Overall	475	20.42	3.36	21.11/226	19.80/249

ANCOVA 1996 (n 495)	*Covariate (NI En) F = 179.01, df 1, significance <0.001* *Gender F = 3.73, df 1, significance <0.05* *Question Choice F = 1.83, df 3, n.s.* *Interaction of Gender & Q. Choice F = 0.92, df 3, n.s.*
ANCOVA 2000 (n 475)	*Covariate (NI En) F = 200.79, df 1, significance <0.001* *Gender F = 18.96, df 1, significance <0.001* *Question Choice F = 1.04, df 3, n.s.* *Interaction of Gender & Q. Choice F = 1.25, df 3, n.s.*

As was the case when the 1996 version was contrasted experimentally with the 1999 version, the first option on the 1996 test (a narrative) was by far the most popular and the last (an informative writing task) proved least popular. The relative popularity of the four options in the 2000 version was far more even. Option 3, the first of the two narrative writing options featured, attracted only a few more takers than the two tasks which preceded it, and whilst the

last option was the least popular it nonetheless attracted a substantial proportion of the children. Option 1 (a diary) attracted many more girls than boys, who were more likely to choose option 3.

An important question here is whether or not the choice between options makes any difference to the mark children are awarded. The ANCOVA analyses indicate that, after controlling for any variations in the calibre of the groups selecting the various tasks, differences between the marks awarded to the groups selecting each option within both the 1996 and 2000 versions proved statistically insignificant. This replicates the findings when the 1996 version was contrasted with the 1999 version. The optional questions do not seem to have varied in 'difficulty' - which might have arisen from variations in either the accessibility of the tasks or the marking process. Children have not been significantly advantaged or disadvantaged by their choices.

But our central concern with differences between versions brings us back to the differences between average Writing marks obtained by those taking the different versions of the test.

The detection of statistical significance is a product of sample sizes as well as the size of the effects observed. These analyses of option effects inevitably have less 'power' to detect significant differences because the sub-groups are smaller, especially when one option attracts most of the children, as with the 1996 version.

Leaving 'significance' aside, what are the implications of the variations in mean marks? It was notable that of the 1996 options, that with least scaffolding of stimuli, prompts and planning support materials (3 - The longest day) produced the lowest mean mark in both our experimental contrasts. Similarly, in the 1999 version, option 4 had least in the way of supporting materials and the lowest mean mark. Also notable was the pattern of marks in the 2000 version, where four very well supported questions all produced high and very similar means. Clearly this has implications from both learning and measurement perspectives.

Does more extensive scaffolding in writing tasks raise the quality of work produced - and hence the marks/levels achieved? These data hint that it might, but Green (2001) has explored this issue by systematic manipulation of the features of writing tasks and her work suggests that it is not necessarily so. Further empirical research investigating this issue may well be merited.

This also raises interesting questions about the status of 'improvements' in marks generated by the question setter's success in tailoring the questions to elicit desirable responses. Should this be taken into account when setting thresholds? We will return to this matter in section 4 below, in an effort to explore some of the many issues involved.

How might standards in 1996 and 2000 differ?

Table 2.4.2.7 shows the cut-scores determining Levels in the 1996 and 2000 versions of KS2 English. The thresholds governing Levels 2 and 3 in the 2000 version are one mark below those set in the 1996 version, but the 2000 Level 4 and 5 thresholds are eight and nine marks below their 1996 equivalents.

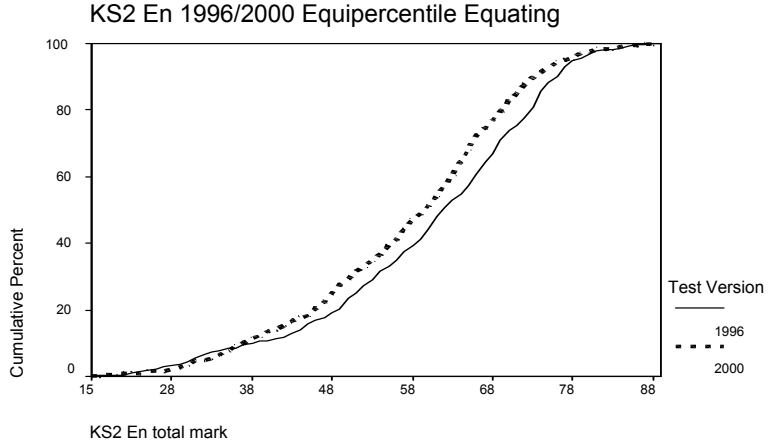
We have already shown that although the 2000 version of the test proved more difficult in this experiment (because the Reading and Spelling components were more challenging) the Levels achieved were higher than on the 1996 version. Cut-scores must change each year to reflect the relative difficulty of different versions of the test, but this variation in the cut-scores would seem to be larger than is merited. How might the mark scales from these two versions of the test be better aligned?

The final column of table 2.4.2.7 shows the outcome of an equipercentile equating, which is also illustrated by figure 2.4.2.2, by providing the 2000 mark equated to each of the 1996 cut-scores.

Table 2.4.2.7 KS2 English cut-scores and equatings (n 970)

KS2 English	*1996 Cut-Score*	*2000 Cut-Score*	*2000 mark Equipercentile equated to 1996 Cut-Score*
NCL 2	27	26	29
NCL 3	30	29	31
NCL 4	57	49	54
NCL 5	79	70	77

Figure 2.4.2.2 KS2 English: Equipercentile equating between 1996 and 2000



The equatings suggests that to align 2000 standards with those set in 1996 would require cut-scores substantially above those chosen. The 2000 cut-cores would need to be raised by three marks at Level 2, two marks at Level 3, five marks at Level 4 and seven marks at level 5. This is very similar to the outcome of our previous experimental equatings between the 1996 and 1999 versions of the KS2 English.

Given that the group assigned to the 1996 version in this experiment may have been marginally superior to those assigned to the 2000 version, it is possible that if anything this analysis slightly underestimates the difference in standards. The gap between the mean levels achieved via the 1996 and 2000 versions reported in table 2.4.2.4 is slightly greater than that in the 1996 v 1999 experimental comparison.

Thus the disparity identified between 1996 and 1999 KS2 English test standards is replicated in the 2000 version of the test. Indeed the difference may have widened slightly, especially at the Level 5 threshold - where national results improved in 2000.

Where and how might standards have diverged?

What else can we say about the source of the discrepancy in the standards set in 1996 and 2000 KS2 English? As was the case when the 1996 version was compared with the 1999 version, we cannot say when the disparities arose. Decisions in any or all of the intervening years - 1997, 1998 1999 and 2000, may have played a part.

We are also, as before, hampered by the absence of separate cut-scores for Reading and (aggregated) Writing in 1996. But as in the contrast between the 1996 and 1999 versions, it is perhaps fortuitous that the aggregated Writing marks on the 1996 and 2000 versions are again very similar here. The higher Writing mark obtained on the 2000 version again largely counterbalances this version's more difficult Spelling test and lower Handwriting marks. In this replication, the mean aggregate Writing marks (formed by the combination of Writing, Spelling and Handwriting) were 30.8 for the 1996 version and 30.3 for the 2000 version.

Given that the aggregated Writing mean marks obtained by the groups assigned to the two versions are so similar, the disparity in standards must therefore, again, arise largely from the Reading component. In this replication, average marks gained by the matched groups on the Reading component were about two marks lower on the 2000 version than on the 1996 version. But in 2000 the cut-scores were eight marks lower at Level 4 and nine marks lower at Level 5. Thus the disparity in standards implied by the adjustments to cut-scores required for equivalence suggested above must stem largely from the thresholds set for the Reading component.

This has important implications for the interpretation of changes in the pattern of KS2 English results at a national level over recent years. There appears to be a widespread impression that in recent years children's reading has improved whilst their writing has not. This view was apparently shared by the Secretary of State, who announced initiatives to improve the teaching of writing at the time the initial results of 2000 national assessments were released.

But the combined evidence from our experimental contrasts between the 1996 version and the 1999 and 2000 versions, suggests that some of the recent improvements in Reading results (especially at and above the expected Level (4)) are illusory, arising from the different national tests and associated level thresholds used in different years.

Conversely the standards for marking Writing seem to have been maintained, although further research may be merited, designed to see if some of the small improvements in marks/levels noted over time may relate to the development of writing tasks which make it easier for children to produce writing with the desired characteristics.

We should also recognise the danger that the aggregation of Spelling with Writing may mask shifts in writing achievement. Spelling tests are, perhaps inherently, more likely to vary in difficulty between years than the other Writing (and Handwriting) assessments.

2.5 KS2 Mathematics experimental comparison: 1996 v 1999

Historical Trends

The percentages of children in England achieving each level in the KS2 Mathematics tests set from 1996 to 1999 are shown in table 2.5.1., where it is apparent that results have improved quite dramatically since 1996, when a total of 54% of the age cohort reached or exceeded level 4. By 1999 this level was reached by 69% of children, substantial progress having been recorded in 1997 and again in 1999 itself.

Table 2.5.1 Key Stage 2 Maths 1996-1999 (% at each level*)

Year	*Below 3*	*Level 3*	*Level 4*	*Level 5*
1996	8%	34%	40%	14%
1997	7%	28%	44%	18%
1998	7%	31%	42%	17%
1999	6%	23%	45%	24%

* Rows do not total 100% as absentees and children who were disapplied are not included.

The validity of experimental comparisons in schools in Northern Ireland

Variations in style and content between the 1996 and 1999 versions

The 1996 and 1999 versions of the KS2 Mathematics test both included two broadly similar 45 minute written Papers (A and B). But there was a notable difference between these two versions because of the presence of the Mental Arithmetic test (based on a 20 minute audio-tape) in the 1999 version. This element, carrying 20% of total marks in 1999, was first introduced in 1998.

The range and balance between item types in the written papers was similar in the 1996 and 1999, except that fewer marks were attached to questions demanding an explanation in the 1999 version.

The overall balance of marks available between different attainment targets remained broadly similar, although the marks available for Data Handling were slightly lower for the 1999 version than in 1996, whilst Number received greater emphasis in 1999.

At a greater level of detail, there were changes of emphasis on some mathematical content between the two tests. Probability questions were noticeably more prominent in 1999 than 1996, as were questions relating to Measures. Negative Numbers also featured in 1999 whilst being absent in 1996. To balance these emphases, the 1999 test contained fewer questions on Data Handling and Fractions compared to the 1996 version.

Structural and curricular differences between England and Northern Ireland

Project staff undertook a detailed desk review of the official documentation available describing primary schools and the Mathematics curriculum in

England and Northern Ireland, to identify features which might affect the tests from 1996 and 1999 unevenly and so bias our experimental findings.

Structural issues
In England KS1 spans ages 4-7 years – the reception class and years 1 and 2 (Y0 - Y2): whilst in Northern Ireland it extends to include ages 4-8 years – Primary 1 to 4 (P1 - P4). In both systems KS2 ends at age 11, spanning Y3, Y4, Y5 and Y6 in England and P5, P6 and P7 in Northern Ireland. Thus the NI children in P7 who took part in the experimental KS2 Mathematics testing had reached the end of 'their' KS2, having enjoyed a similar length of schooling to English children at the same stage.

In Northern Ireland statutory assessments at the end of KS2 are by teacher assessment, supported by the use of a selection from a catalogue of assessment units distributed by NICCEA, who also moderate schools' assessment portfolios. This NI system uses national curriculum 'levels' like those employed in England, awarded on a 'best-fit' basis. Target setting and benchmarked comparisons of schools' achievements in national tests were introduced in England in 1998, but were not a feature of primary education in NI.

Although NI KS2 children do not experience externally set and marked end of key stage tests like those set in England, many of them do encounter similar high stakes testing in another form. At the end of KS2, NI children wishing to transfer to selective grammar schools take (optional) transfer tests in English, Mathematics and Science which are externally set and marked. In 1998/9 these transfer tests were taken by 67% of 11 year olds, so a majority of children in NI were likely to be familiar with formal test conditions. Even if some are less test-wise than others, the design of our study should ensure that random groups are assigned to different versions, so that comparisons between them remain valid. There seems no reason why these differences in assessment arrangements should bias our comparisons between the 1996 and 1999 test forms.

Curricular issues
In both England and NI the curriculum was defined within programmes of study and attainment targets. For NI, expected levels of performance were set out in level descriptions for each attainment target. Their KS2 programmes of study extend to level 5 and it is expected that the majority in NI will be working at either level 3 or level 4 at the end of KS2. The majority of English children are expected to be working at level 4 at this stage.

The NI KS2 Mathematics curriculum has five attainment targets compared to England's four, because NI's ATs 3 & 4 (Measures & Shape and Space) correspond to AT3 in England. In both curricula, attainment targets relate directly to sections within the programmes of study, which (with some exceptions) cover much the same subject content. Some things (e.g. experience with calculators) are found in the introduction to the NI programme of study rather than the programme itself. By and large, level descriptors

correspond too, although there are instances where items appear at different levels.

Emphasis on a few topics differs enough between the curriculum documents for England and NI to raise some concern that it might affect children's performance in the tests being compared - if these variations in curriculum documentation do cause teachers in NI or England to vary the emphasis within their teaching significantly. We must recognise that we have little evidence of the extent that such variations in emphasis do exist in practice[1]. The topics our review identified as potentially problematic in this respect and which feature in the 1996 or 1999 tests are described below.

Shape and Space and Measure were accorded greater detail in the NI curriculum than in England and children from NI might thus have an advantage on some such topics. Conversely England's national curriculum detailed various topics, including aspects of probability; the calculation of fractions or percentages of quantities; rotational symmetry and aspects of co-ordinates, which did not feature in the NI KS2 curriculum documents. Again, teachers may still include these in their teaching programmes, curriculum documents notwithstanding.

The overall impression from these analyses suggested that on each version there were a few questions where NI children might have some potential advantage. Balancing these were a few others where the boot might be on the other foot and NI children might be at a disadvantage compared to those from England. But the balance of relative advantage seemed very consistent in 1996 and 1999, suggesting that it would affect both years' tests in much the same way and hence not invalidate our comparisons.

Teachers' opinions

Given the difficulty of knowing how the variations in curriculum documents might actually affect what teachers do, the opinions expressed by teachers from the schools taking part in this experiment are of great interest. In this phase of the Project fieldwork was undertaken by project staff visiting schools to provide feedback: using a structured interview schedule which included (amongst others) questions asking if pupils' learning had left them under-prepared or exceptionally well prepared, for particular questions.

Several teachers commented on similarities between the KS2 Mathematics tests and Northern Ireland's 11+ Transfer tests, and the QCA tests were considered well matched to Northern Ireland's curriculum. Reactions to the mental arithmetic test were favourable. Teachers liked the use of a tape recording and the prompts on the children's answer sheet.

Some schools thought that their children might not be familiar with calculator work, whilst others suggested that their children might not be used to showing explanations or reasoning as part of a written answer.

[1] Plewis and Veltman (1996) collected data suggesting that the introduction of the national curriculum may have reduced variations in curricular coverage in mathematics at KS1 in Inner London, and which 'support a fair degree of consistency between the official and taught curriculum'.

Although a few teachers mentioned two questions in the 1996 test (Q14 and Q18 in Test A) which might have been unfamiliar to their children, in general both versions seem to be regarded as fair tests.

In summary
This review suggested that there may be topics absent or emphasised less in one system or another but that test items likely to be affected were distributed between the 1996 and 1999 tests in such a way that it should not affect experimental comparisons.

The teachers consulted via fieldwork seemed confident that both years' tests were valid for their children and it seems reasonable to accept their judgement and regard the experiment as a fair comparison.

The data

A random sample of schools (excluding schools with cohorts below 15, for logistical reasons) were approached (on the project's behalf) by NICCEA and asked to participate. In total 22 of the schools asked to administer the KS2 Mathematics test (with a total of 1,168 P7 children on roll) agreed to do so.

Test materials and full instructions for test administration (versions of QCA's instructions, edited by project staff to facilitate simultaneous administration of the two versions) were supplied to the schools, including arrangements for the spiral allocation of random groups of children to the 1996 and 1999 versions of the test. Because there was no mental arithmetic element in the 1996 version we asked all children in this experiment, including those assigned to the 1996 version, to take the 1999 mental arithmetic test - although, naturally, this was only used in calculating test levels for the group assigned to the 1999 version. Testing took place in late May 1999, soon after operational testing was finished in England. All these schools returned completed tests to UCLES in Cambridge, together with details of each child's gender and their end of KS2 assessments in Mathematics; the latter conducted according to NICCEA's instructions and subject to their monitoring arrangements.

The KS2 Mathematics tests were marked by a team of 2 suitably experienced KS2 markers - including a team leader with experience of marking KS2 Mathematics throughout the period 1996 -1999. Scripts (for both the 1996 and 1999 versions) from each school were randomly apportioned amongst the markers, to ensure that inter-marker variation could not be confounded with test form. Marking took place at UCLES Research & Evaluation Division's offices, supervised by project staff with primary experience. The operational marking schemes for 1996 and 1999 were followed rigorously throughout and statistical monitoring of the marks awarded revealed that variations in the marks awarded by different markers were insignificant, indicating that marking was conducted fairly and accurately for both versions of the test.

Following data entry, total marks were computed and checked against markers' totals and differences were reconciled. Further data checks included the numbers allocated to each version and mean scores within schools, to establish that these were consistent with random allocation to test forms. In a small number of cases this appeared to be in doubt and data from these teaching groups were excluded from subsequent analyses.

The analyses reported below are based on the 945 children for whom full sets of data were available, including their performance on the 1996 or 1999 versions of the KS2 Mathematics tests, their gender and their Northern Ireland KS2 Mathematics assessments.

How do results on the 1996 and 1999 test forms compare?

Before we can answer this question we have to check if the groups assigned to take the two versions of the KS2 Mathematics test are of equivalent ability. Table 2.5.2 shows how the groups assigned to the 1996 and 1999 versions of QCA's KS2 Mathematics tests appeared to be well matched - in terms of their NI KS2 mathematics assessments. Random assignment appears to have worked reasonably well, as although the group assigned to the 1999 version included a slightly higher proportion reaching NI NCL 4 or NI NCL 5, the variations were no greater than might be expected by chance.

Table 2.5.2 NI KS2 Mathematics Assessment Levels by test version

	NI NCL 1	*NI NCL 2*	*NI NCL 3*	*NI NCL 4*	*NI NCL 5*		
1996	0.2	2.9	22.0	36.1	38.8	100%	n 487
1999	0.2	3.5	16.6	36.9	42.8	100%	n 458
n 945	*Chi-square Likelihood Ratio 4.79, df 4, n.s.*						

Table 2.5.3 shows that whilst the group assigned to the 1996 version included rather more boys than girls, the reverse was true for the group taking the 1999 version. Again the difference in this respect is not statistically significant, but because of the possibility that there may well be sex differences in performance in mathematics it may be prudent to consider gender in making comparisons between versions.

Table 2.5.3 Test version by gender

	boys	*girls*		
1996	52.6	47.4	100%	n 487
1999	48.7	51.3	100%	n 458
n 945	*Chi-square Likelihood Ratio 1.42, df 1, n.s.*			

Table 2.5.4 shows the percentages, overall, achieving each level via the 1996 and 1999 versions of KS2 Mathematics. From this it would appear that the group assigned to the 1999 version of the test obtained slightly better results than those taking the 1996 version, with 85.4% obtaining NCLs 4 or 5 on the 1999 version, compared with 79.3% on the 1996 test.

Table 2.5.4 KS2 Mathematics Levels by test version

	NCL <2	NCL 2	NCL 3	NCL 4	NCL 5		
1996	1.2	0.8	18.7	41.9	37.4	100%	n 487
1999	3.1	0.4	11.1	45.0	40.4	100%	n 458
n 945	*Chi-square Likelihood Ratio 14.54, df 4, significance <0.01*						

But we have already suggested that it would be prudent to bear sex differences in mind. Tables 2.5.5a and 2.5.5b respectively show the Northern Ireland Mathematics assessments and the QCA KS2 Mathematics NCLs gained by girls assigned to the two versions of the test. Equivalent information for boys is presented in table 2.5.6a and 2.5.6b.

Table 2.5.5a Girls Northern Ireland KS2 Mathematics Levels by test version

	NI NCL 1	NI NCL 2	NI NCL 3	NI NCL 4	NI NCL 5		
1996	0.4	2.3	16.8	39.5	41.0	100%	n 256
1999	0.4	2.7	14.3	40.8	41.7	100%	n 223
n 479	*Chi-square Likelihood Ratio 0.59, df 4, n.s.*						

Table 2.5.5b Girls KS2 Mathematics Levels by test version

	NCL<2	NCL 2	NCL 3	NCL 4	NCL 5		
1996	0.8	1.2	14.5	43.8	39.8	100%	n 256
1999	2.7	0.4	10.8	43.5	42.6	100%	n 223
n 479	*Chi-square Likelihood Ratio 4.98, df 4, n.s.*						

The NI Mathematics assessments show that the groups of girls assigned to the two versions were clearly very alike in mathematics achievement and the results they obtained on the 1996 or 1999 versions of KS2 Mathematics were also remarkably similar; providing no reason to suggest that standards over time might be adrift.

Table 2.5.6a Boys Northern Ireland KS2 Mathematics Levels by test version

	NI NCL 2	NI NCL 3	NI NCL 4	NI NCL 5		
1996	3.5	27.7	32.5	36.4	100%	n 231
1999	4.3	18.7	33.2	43.8	100%	n 235
n 466	*Chi-square Likelihood Ratio 5.91, df 3, n.s.*					

Table 2.5.6b Boys KS2 Mathematics Levels by test version

	NCL<2	NCL 2	NCL 3	NCL 4	NCL 5		
1996	1.7	0.4	23.4	39.8	34.6	100%	n 231
1999	3.4	0.4	11.5	46.4	38.3	100%	n 235
n 466	*Chi-square Likelihood Ratio 12.53, df 4, significance 0.01*						

The groups of boys taking the two versions were less well matched, with those taking the 1999 version having better NI KS2 assessments (77% reaching level NI NCL 4) than those taking the 1996 test (68.9% reaching NI NCL 4). However, the variation between distributions of NI Mathematics assessments did not reach statistical significance. KS2 test results were likewise better for the group taking the 1999 version and the difference here did exceed the levels we could expect to arise by chance. We must however consider the possibility that these differences arose because they were more able, as their NI Mathematics assessments hinted, rather than from variations in test standards.

To do this we assumed that KS2 (& NI) national Curriculum Levels enjoy the properties of an equal interval scale. We could then summarise the levels

achieved by boys and girls taking each version by the means and standard deviations of their levels and carry out analyses of variance (ANOVA) and analyses of covariance (ANCOVA) to test the significance of gender and test version effects. The results are presented in table 2.5.7

These confirmed that, on average, girls achieved higher levels than boys on both versions of the test, but the differences observed were quite small, less than one fifth of a level on the 1996 version and less than one tenth of a level on the 1999 test.

Overall, the difference in average levels obtained by the groups assigned to the two versions was only 0.04 of a level (averages for girls taking the two versions are the same) and ANOVA suggested that the differences between groups taking the two test versions were not statistically significant, whilst gender differences bordered on those we might expect by chance. An ANCOVA analysis, controlling for children's achievement levels in Mathematics via their NI KS2 Mathematics assessments, confirmed that differences between the levels achieved via the 1996 and 1999 versions of the test were not significant.

Table 2.5.7 KS2 Mathematics mean transformed NCLs by gender and test version

		Mean	*SD*	*n*
1996		4.12	0.88	487
	boys	4.03	0.94	231
	girls	4.20	0.82	256
1999		4.16	1.00	458
	boys	4.12	1.02	235
	girls	4.20	0.97	223
Total		4.14	0.94	945

ANOVA	*Gender F = 3.94, df 1, significance 0.05*
(n 945)	*Test Version F = 0.56, df 1, n.s.*
	Interaction of Gender & Version F = 0.44, df 1, n.s.
ANCOVA	*NI Mathematics NCL F = 1,253.06, df 1, significance <0.001*
(n 945)	*Gender F = 0.77, df 1, n.s.*
	Test Version F = 0.41, df 1, n.s.
	Interaction of Gender & Version F = 0.06, df 1, n.s.

So in this case the null hypothesis stands. The differences between the groups assigned to the two versions of the test were not significant when we had controlled for variations in achievement in NI KS2 Mathematics assessments, so there was no suggestion of any difference between the standards of the KS2 Mathematics tests from 1996 and 1999.

How did children perform on the components in each year's test?

Table 2.5.8 gives the means and standard deviations of scores on each component in the 1996 and 1999 versions of the test, including those for each gender.

Girls' mean scores were higher than those for boys in both Test A and Test B in both versions, with boys' scores displaying slightly greater dispersion, as is often found elsewhere (Gipps & Murphy, 1994; Johnson, 1996). This pattern was broken for mental arithmetic, where boys and girls obtained very similar mean scores.

Table 2.5.8 Means and standard deviations of scores on components by gender

		All	*n = 487*	*Girls*	*n = 256*	*Boys*	*n = 231*
		mean	*sd*	*mean*	*sd*	*mean*	*sd*
1996	*Test A*	26.84	7.64	27.45	7.47	26.17	7.78
	Test B	26.16	7.91	26.68	7.32	25.36	8.47
	Mental Arithmetic (1999)[1]	15.58	4.10	15.65	4.11	15.50	4.09
	[1]n 475–250 girls/225 boys						
		All	*n = 458*	*Girls*	*n = 223*	*Boys*	*n = 235*
		mean	*sd*	*mean*	*sd*	*mean*	*sd*
1999	*Test A*	27.73	7.75	27.95	7.40	27.52	8.08
	Test B	26.90	8.76	27.64	8.30	26.19	9.15
	Mental Arithmetic	15.95	4.06	15.93	3.95	15.97	4.16

Mean scores on the 1999 versions of both Test A and Test B were marginally higher than scores on the 1996 versions. The mean score obtained by the group assigned to the 1999 version of the test on the 1999 Mental Arithmetic test (taken by both groups) was also marginally higher than that of the group assigned to the 1996 version.

This pattern is entirely consistent with the evidence (discussed above) that the 1999 group were of a slightly higher calibre, thus meriting the rather better levels they achieved.

Again there is no suggestion here that standards in the 1996 and 1999 versions of the KS2 Mathematics test might vary.

2.6 KS2 Science experimental comparison: 1996 v 2001

Historical Trends

The percentages of children achieving each level in successive years between 1996 and 2001, shown in table 2.6.1, suggests a massive improvement in teaching and learning in KS2 Science, the proportion achieving levels 4 and 5 having grown from 62% in 1996 to 87% in 2001. After slipping back slightly in 1998 results improved remarkably in 1999 and have continued to advance since then.

Table 2.6.1 Key Stage 2 Science 1996-2001 (% at each level*)

Year	*Below 3*	*Level 3*	*Level 4*	*Level 5*
1996	6%	28%	48%	14%
1997	4%	23%	50%	18%
1998	4%	23%	53%	16%
1999	3%	16%	51%	27%
2000	3%	11%	50%	34%
2001	1%	9%	53%	34%

* rows do not total 100% as absentees and children disapplied are not shown

The validity of experimental comparisons in schools in Northern Ireland

Variations in style and content between the 1996 and 2000 versions
Both the 1996 and 2001 versions of the KS2 Science tests contain two papers, A and B. In both versions each paper is thirty five minutes long and carries a maximum of 40 marks.

Question style
Each item was classified (by project staff with suitable expertise) according to the type of item. The marks available for the various types of items encountered are shown in table 2.6.2.

Table 2.6.2 Marks available to different item types

*Item type**	*Objective*	*Diagram*	*One Word*	*Short Answer*	*Explain*
1996 version (marks available)	29	8	14	19	10
2001 version (marks available)	29	4	25	18	4

* Objective items involve multiple choice questions or ticking boxes or matching given information etc.
Diagram questions require drawing or adding labels etc. to a diagram.
One word items require one word answers.
Short answer items require brief phrase(s) or sentence(s).
Explanation items require longer written answers and the command word in the item is 'explain'.

A greater number of 'one word' questions were found in the 2001 version. Overall, eleven more marks were available for this type of question in the 2001 version of the test, mainly at the expense of diagrammatic and, most notably, explanation questions; which both carried proportionately more marks in the 1996 version of the test.

The 2001 version of the test was also longer - at 35 pages, containing 239 sentences / 2831 'words', compared to 30 pages with 153 sentences / 1778 'words' in the 1996 version.

Content
A further classification, identifying the topics and hence the national curriculum attainment targets (ATs) tested by each item, forms the basis for table 2.6.3. This shows the balance of marks available for each AT in each version of the test. In the curriculum changes for 2000 for England, aspects of science previously set out in the introduction to the Programmes of Study have been incorporated into Sc1 (Scientific Enquiry) and both the 1996 and 2000 versions were classified on this basis. The curriculum changes in 2000 also removed two topics from KS2 Science: 'saturated solutions' and 'balanced/unbalanced forces'.

Table 2.6.3 Marks available for each Attainment Target.

*Attainment Target**	*AT1*	*AT2*	*AT3*	*AT4*
1996 version	11	25	23	21
2001 version	14	24	20	22

* AT1 Scientific Enquiry
AT2 Life Processes and Living Things
AT3 Materials and their Properties
AT4 Physical Properties

In table 2.6.3, it can be seen that on the whole, no substantial differences in the balance of items between ATs were observed between the two versions, although slightly more questions (and hence marks), assessing AT1 skills were available in the 2001 version of the test. This was mainly at the expense of questions testing AT3.

At a greater level of detail, within each attainment target, there were differences in mark allocations between versions for some of the topics assessed. These were not usually large and the more notable variations observed are listed below:

- AT3 changing materials (11 marks in 1996 : 8 marks in 2001)
- AT2 green plants (6 marks in 1996: 2 marks in 2001)
- AT4 electricity (5 marks in 1996 : 1 mark in 2001)
- AT4 forces & motion (6 marks in 1996: 10 marks in 2001)

These variations in the style and content of the tests do not in themselves appear to threaten the validity of our experimental comparisons between versions, although the reduced number of marks available in 2001 for questions requiring children to 'explain' might have a bearing on the outcome.

Curricular and structural issues
Structural variations in the organisation of schools in Northern Ireland have been discussed earlier and would not seem to threaten the validity of our experimental comparisons at KS2 in science.

Curricular variations

The programmes of study for science in England and NI in general cover similar areas and expected outcomes, although minor variations in terms used and in the topics required to be covered do exist. Overall, level descriptions cover much the same criteria, although some items appear at different levels. However, aspects of the English science curriculum which do not feature explicitly in the NI curriculum are:

AT2: Life Processes and Living Things

- Life processes common to humans and animals (nutrition/reproduction).
- Life processes common to plants
- Teeth
- Function of root/feeding
- Micro-organisms

AT3: Materials and their Properties

- Separating mixtures of materials

AT4: Physical Properties

- Circuit diagrams
- Magnets/attraction & repulsion
- Measuring forces
- Light enters the eye - seeing
- Pitch and loudness
- The earth & beyond

The questions affected by these differences and the resulting number of marks affected are shown in table 2.6.4 below.

Table 2.6.4 Questions affected by curricular disparities between England & NI

1996 Paper A marks	*1996 Paper B marks*	*2001 Paper A marks*	*2001 Paper B Marks*
Q2b) Magnets (attract/repel) ….2 marks *Q3aii)* Light entering the eye……1 mark *Q4d)* Life processes (humans & animals) ……3 marks *Q5e)* Measuring forces……..1 mark *Q9a)* Pitch……..1 mark *Total 8/40 marks*	*Q2d)* Roots…….2 marks *Q3)* Teeth…….4 marks *Q4a/c/d/e)* Sun/shadow ……4 marks *Q8c)* Circuit diagrams ……2 marks *Total 12/40 marks*	*Q1d)* magnets (attract/repel)…..1 mark *Q2a/b/c)* Teeth...3 marks *Q4c)* Light entering the eye……1 mark *Q6d)* Roots……..1 mark *Q8b)* Life processes (humans & animals)… ……1 mark *Q9d)* Separating materials……...1 mark *Total 8/40 marks*	*Q2d)* Life processes (plants nutrition)..1 mark *Q6a-d)* Earth & beyond …….4 marks *Q9c)* Circuit diagrams ……1 mark *Q11c)* Loudness….. ……2 marks *Total 8/40 marks*

Curriculum documentation notwithstanding, children in NI might encounter such topics if primary science teachers there include them in their plans for teaching and learning. The Project's survey of reactions to the test materials by teachers in participating schools (reported below) investigated this possibility.

Teachers' opinions

Fifteen of the eighteen primary schools involved in the experiment returned questionnaires seeking their views. At a summary level, schools were asked if the fit of the 1996 and 2001 versions of the KS2 Science test to the NI curriculum was 'very close', 'similar' or 'not at all'. Table 2.6.5 reveals that

none thought the match worse than similar, although more thought the 2001 version a very close match than the 1996 version.

Table 2.6.5 Fit of test versions to NI curriculum (n 15)

	1996 version	*2001 version*
very closely	1	7
similar	11	6
not at all	0	0
no response	3	2

General comments (from a total of seven schools) confirmed that the tests were by no means unsuitable for children from NI. All seven made positive comments remarking on the quality of the structure and presentation of the tests. However several described the tests as 'challenging', suggesting that the children were required to answer in more depth than would be required in the Northern Ireland transfer tests. Two schools also noted that they felt the tests demanded not only a high standard of scientific knowledge but also high quality literacy skills and language development.

For both versions of the test, however, schools reported specific disparities between the tests and the Programme of Study in Northern Ireland. Certain scientific concepts and vocabulary from the tests were repeatedly cited as aspects of science not studied in NI, or alternatively, as unfamiliar terms for their pupils[1]. These were:

1996 test

- 'magnetism' (7 of the 15 schools noted this disparity)
- 'circuit diagrams' (4 schools)
- 'contraction of human muscles' (4 schools)
- 'pitch' (4 schools)
- 'food chain vocabulary (4 schools)
- 'movement of the sun' (3 schools)
- 'gravity' (3 schools)

2001 test

- 'magnetism' (9 schools)
- 'earth and beyond' (7 schools)
- 'circuit diagrams' (3 schools)
- 'forcemeter' (3 schools)

Except for the reference to 'human muscle use', all of the above (amongst others) were amongst those identified by project staff (above) as potential differences between the English and Northern Ireland programmes of study in Science.

The teachers were then asked to pinpoint specific questions from the tests that they felt such curricular disparities might affect[2]. The following list shows questions identified in each version of the test and the number of schools[3] doing so.

[1] Concepts that were identified by 3 or more schools are listed.
[2] Note that 2 schools did not complete this section of the questionnaire.
[3] Questions that were identified by 3 or more schools are listed.

1996 test

- Paper A Q2 (magnets) 8
- Paper A Q4 (food chain vocabulary unfamiliar) 7
- Paper A Q5 (forcemeter) 4
- Paper A Q9 (pitch) 3
- Paper A Q10 (muscles) 5
- Paper A Q11 (complicated graph) 3
- Paper B Q4 (movement of the sun) 3
- Paper B Q8 (circuit diagrams) 5

2001 test

- Paper A Q1 (magnets) 9
- Paper B Q3 (forcemeter) 5
- Paper B Q6 (earth and beyond) 10
- Paper B Q9 (circuit diagrams) 6

Once again, the questions identified above match those previously identified by project staff as potentially problematic for children in Northern Ireland. The exceptions here are again, the 1996 Paper A question on muscles (Q10), 1996 Paper A question 11 and 2001 Paper B question 3.

With respect to the validity of our comparisons, the questionnaire data indicates that 15 marks in the 1996 test could be affected (13 of these marks match with those identified in the Project's desk analyses), compared with 7 marks from the 2001 version (where 6 marks match desk analyses). The suggestion here is not that these marks will be inaccessible to all children in NI. Some will have learned these things and will complete these questions successfully, as observed performances in the test confirm. Rather, it is arguable that compared to children in English schools they are less likely to succeed. The chief problem for our comparison between versions is that the effects may apply unevenly between the 1996 and 2001 test forms because more items are affected in one than the other. This evidence therefore supports, and even enhances, the desk analysis' indication that scores on the 1996 version were, potentially, more likely to be depressed by children not having encountered some of the topics involved. This might have the effect of biasing the experiment; putting the null hypothesis into some doubt. The potential direction of the effect is however clear and it can be considered when interpreting the outcomes observed.

The Data

Eighteen primary schools agreed to assist with the Project and administered the two versions of the test to their P7 (as the final year of KS2 is known in NI) children in June 2001. Schools also provided their statutory end of key stage assessments in Mathematics (there being no such KS2 assessments in science in Northern Ireland), using a ten level scale like England's as instructed (and moderated) by NICCEA. Test administration clearly went well, despite the complexities of spiral allocation required to create random groups taking the 1996 and 2001 versions of the test. Questionnaire replies from 10 schools (of the 15 responding) reported that they encountered no problems and four others commented on the clarity of the instructions received. The

remaining school noted that their children had struggled to complete the tests within the allocated time limit.

All scripts were returned to UCLES' Cambridge offices, where they were apportioned at random to be marked by a team of four experienced KS2 test markers, led by a senior marker with experience of operational marking for both the 1996 and 2001 versions. Marker co-ordination meetings were held to brief the team, making use of the example scripts used operationally. Data entry was followed by range and totalling checks and investigation of the numbers assigned to each form in each school, together with schools' means and variances, confirming that spiral allocation appeared to have been performed correctly. Analyses also confirmed that any differences between marks awarded by the four markers were within the range to be expected by chance.

KS2 Science test levels were then computed, using the operational cut-scores for either the 1996 or 2001 versions, as appropriate.

Inevitably data for some of the 1000 pupils who participated were incomplete, due to absence for one or other paper or missing NI mathematics assessments. Complete data were obtained for 952 children, forming the basis for the analyses reported below. Of these 256 boys and 230 girls took the 1996 version and 253 boys and 213 girls took the 2001 version.

How do results on the 1996 and 2001 test forms compare?

We need first to establish if the groups assigned to the two versions provide a fair basis for comparisons. Table 2.6.6 shows that the distributions of NI KS2 Mathematics levels for the two groups are remarkably similar, indicating that this is likely to be the case.

Table 2.6.6 NI KS2 Mathematics Assessment Levels by QCA test version

	NI Ma L<2	*NI Ma L2*	*NI Ma L3*	*NI Ma L4*	*NI Ma L5*		
1996	0%	1.2%	17.7%	34.4%	46.7%	100%	n 486
2001	0%	2.1%	17.6%	35.2%	45.1%	100%	n 466

n 952 Chi-square Likelihood Ratio 1.37, df 3, n.s.

Table 2.6.7 QCA KS2 Science Levels by test version

	NCL <2	*NCL 2*	*NCL 3*	*NCL 4*	*NCL 5*		
1996	1.9%	1.4%	27.2%	60.5%	9.1%	100%	n 486
2001	1.3%	0.9%	19.7%	70.4%	7.7%	100%	n 466

n 952 Chi-square Likelihood Ratio 10.86, df 4, significance <0.05

Table 2.6.7 shows that the distributions of KS2 Science test levels achieved on the two versions are distinctly dissimilar, with 78.1% of those taking the 2001 version reaching levels 4 or 5, compared to only 69.6% of those assigned to the 1996 version.

Table 2.6.8 provides more information, showing the means and standard deviations of levels achieved by boys and girls on the two versions of the test, together with an analysis of variance (ANOVA) confirming that the superior

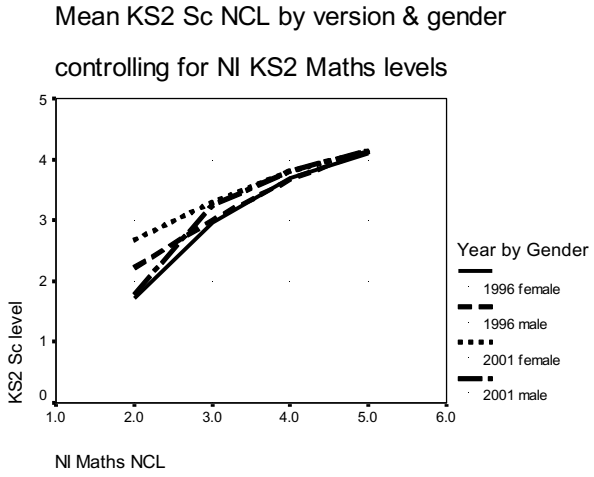
results for those taking the 2001 version appear to be statistically significant. Boys and girls taking the 1996 version obtain very similar results and the ANOVA suggests that the apparently higher average levels achieved by girls taking the 2001 version might arise by chance[4]. This absence of distinctive gender differences in performance (and the significance of the difference in mean levels between test versions) is confirmed by the more powerful analysis of covariance (ANCOVA), also reported in the table. ANCOVA controls for any variations in the abilities of the sub-groups involved by taking NI mathematics assessments into account.

Table 2.6.8 QCA KS2 Science mean NCLs by gender and test version

		Mean	*SD*	*n*
1996	*all*	3.72	0.80	486
	boys	3.71	0.80	256
	girls	3.73	0.80	230
2001	*all*	3.81	0.69	466
	boys	3.76	0.77	253
	girls	3.87	0.59	213
Total		3.76	0.75	952

ANOVA	*Gender F = 1.64, df 1, significance n.s.*
(n 952)	*Test Version F = 4.15, df 1, significance <0.05*
	Interaction of Gender & Version F = 0.79, df 1, n.s.
ANCOVA	*NI Mathematics Assessment F = 535.31, df 1, significance <0.001*
(n 952)	*Gender F = 0.01, df 1, n.s.*
	Test Version F = 8.78, df 1, significance <0.01
	Interaction of Gender & Version F = 0.18, df 1, n.s.

Figure 2.6.1 QCA KS2 Sc Levels by version and gender, controlling for NI Maths



[4] A further ANOVA, of NI maths assessments (not reported here) suggested any variations in KS2 Sc levels by boys and girls on the 2001 version might stem partly from the relatively high ability of girls assigned to this version. Interestingly gender differences in NI Maths were statistically significant, although the corresponding differences in KS2 Science Levels were not.

Figure 2.6.1 displays this information graphically, showing the linear local regressions of the KS2 Science levels achieved by boys and girls assigned to the 1996 and 2001 versions of the test on their NI mathematics assessments.

We should ignore the apparent variations relating to children at level 2 in mathematics as only a handful of the sample are at this level. When children assessed as at levels 3 to 5 in mathematics are considered we can see that boys and girls of similar ability (estimated by maths assessments) taking a given version of the KS2 Science test obtain very similar results, but that those taking the 2001 version have been awarded higher levels (on average).

The differences between versions in levels achieved varies across the ability range. Weaker children (assessed at level 3 in mathematics) on average derived an advantage of 0.33 of a level from allocation to the 2001 version, but the advantage reduced to 0.12 of a level for middle ability children (assessed at level 4 in mathematics), whilst the more able (assessed at level 5 in mathematics) enjoyed little advantage, only 0.04 of a level - all irrespective of gender.

If we take these data as a fair test of the null hypothesis this suggests that the 2001 form of the test confers an advantage over the 1996 form. Our earlier analyses of the tests' content in relation to the NI curriculum, together with NI teachers' opinions, did lead us to question the validity of the null hypothesis, but our suspicion was that curricular variations might make the 1996 version a little less accessible to children in NI than the 2001 version. This directional effect would tend to produce the opposite result to the differences in levels observed here. We can therefore regard our reservations regarding the validity of the experiment as conservative with respect to the outcomes, making us more, rather than less, certain that there is a difference in test standards between the 1996 and 2001 versions of the KS2 Science test.

How might standards in 1996 and 2001 differ?

Table 2.6.9 shows the cut-scores for the award of each level in the 1996 and 2001 versions of the KS2 Science tests. Those for the 2001 version fall five marks below their equivalents for the 1996 version at levels 2 and 3, six marks lower at level 4 but only one mark lower at level 5. But this alone tells us nothing, because variations in relative difficulty between versions might in principle justify such differences - although the results reported above suggest that this is not in fact the case here.

Table 2.6.9 KS2 Science cut-scores and equatings (n 952)

KS2 English	*1996 Cut-Score*	*2001 Cut-Score*	*2001 mark Equipercentile equated to 1996 Cut-Score*
NCL 2	20	15	17
NCL 3	23	18	20
NCL 4	45	39	43
NCL 5	64	65	65

How might the two mark scales be aligned? Table 2.6.9 also shows the result of an equipercentile equating between these two forms of the test, based on the data from the experimental comparisons. Given equivalent groups taking the two versions - a justifiable assumption for our data - equipercentile equating defines marks on the two forms as equivalent if they are achieved by the same cumulative proportion of the children taking the test.

The final column of the table displays the 2001 mark equated to each of the 1996 version's level thresholds. These suggests that to achieve parity with the 1996 version, the 2001 thresholds needed to be increased by two marks at levels 2 and 3, and four marks at level 4, although the level 5 threshold should remain unchanged.

This pattern of adjustments would seem to match the results from the statistical analyses reported earlier. Whilst such adjustments are not large, they would be likely to have made a substantial difference to the pattern of results, nationally, in 2001. But without access to the operational mark distributions it is difficult to estimate accurately how many children would have been awarded lower levels. Adjustments of this magnitude would not however negate more than a modest portion of the huge national gains in achievement since 1996 on KS2 Science tests.

2.7 KS3 English experimental comparisons: 1996 v 2001

The historical pattern of KS3 English test results since 1996 is shown in table 2.7.1 and indicates a noticeable improvement. For instance whilst only 26% obtained level 6 or better in 1996, 31% did so in 2001. The pattern has not been one of steady improvement however. Following a small decline in 1997, 1998 saw results improve markedly, with 35% reaching level 6. But this was not maintained and the percentage reaching level 6 fell back to 28% in both 1999 and 2000. This advance of 5% in the numbers reaching level 6 is modest by comparison with the 'equivalent' improvements in national test results in KS3 Science and KS3 Mathematics, or with KS2 English - where the numbers reaching level 5 more than doubled (12% to 29%) between 1996 and 2000.

Table 2.7.1 KS3 English test results 1996 - 2001 (% of cohort at each level*)

	<4	*Level 4*	*Level 5*	*Level 6*	*Level 7*	*Level 8*	*EP*
1996	15%	23%	31%	18%	7%	1%	0%
1997	12%	27%	34%	17%	5%	1%	0%
1998	12%	19%	30%	25%	9%	1%	0%
1999	12%	20%	36%	21%	6%	1%	0%
2000	11%	21%	35%	21%	6%	1%	0%
2001	12%	20%	33%	22%	8%	1%	0%

* rows do not total 100% as absentees and children disapplied are not shown

The validity of experimental comparisons in schools in Northern Ireland

Variations in style and content between the 1996 and 2001 versions

The structure of the test was largely unchanged between 1996 and 2001. In both versions KS3 English tests had only one tier, involving two papers.

Paper 1
Paper 1 covered levels 4 to 7 and had three sections, A, B and C, which assessed reading and writing, with 61 marks available in total. 1 hour 30 minutes were allowed for children to complete the paper (plus 15 minutes reading time). In 2001 the rubric gave more precise *advice* on the time which should be spent on each question, but there was no means of enforcing this advice and no change to the total amount of time allowed.

In the 1996 version Section A had 2 compulsory questions, carrying 11 and 6 marks respectively, both based on a passage about the first woman to travel alone to the North Pole. Section B had one compulsory question, worth 11 marks and also based on a text - an advertisement for a holiday cruise to Antarctica. Section C had one question (but with a choice between three alternative stimuli / writing genres) carrying 33 marks and testing writing.

In the 2001 version Sections A and B were again based on two texts, and were essentially quite similar in question style to their equivalents in the 1996 version, although this time the shorter of Section A's two questions (based on a newspaper report about the total eclipse of the sun in 1999 and carrying 6

and 11 marks respectively) came first. Section B's 11 mark question was based on an account of the volcanic eruption on Krakatowa in 1883. As in 1996, Section C contained a single question carrying 33 marks testing writing, again with a choice between 3 alternative stimuli. But in 2001 the assessment criteria were stated explicitly on the question paper and, overall, the writing prompts gave more emphasis to purpose and audience.

The mark schemes for questions in Sections A and B are similarly structured, apart from minor differences affecting very few pupils[1]. The (level related) assessment criteria governing the award of marks show many common features, but they are tailored to the different questions/stimuli set in the two versions and include extensive exemplar materials developed during trialling of the tests. The level related assessment criteria governing the award of marks (on a best fit basis) for the writing assignments in Section C are essentially generic in form, with slight variations in wording to make them relevant to the different genre of the three alternative tasks within each version. But the sets of criteria on which assessment criteria in the two versions were based remained unchanged over the period 1996 to 2001. Handwriting and spelling were included in the criteria for each level when assessing writing in both versions.

The style and content of Paper 1 in the versions of the test to be compared is thus very similar. They can reasonably be considered parallel forms.

Paper 2

The same cannot be said of the two versions of Paper 2 (1 hour 15 minutes in both versions), which consisted of questions based on scenes taken from selected Shakespeare plays. Each year three plays were selected and schools could choose which one of these their pupils would study. Each version of the test took two scenes from each selected play as the basis for six alternative tasks, and pupils could choose which one of these tasks to attempt. Thus in practice each pupil could have a choice between two tasks, each based on a different scene from the play s/he would have studied in class in preparation for their statutory end of key stage assessment.

But whilst the form of the test was unchanged between 1996 and 2001, the rotation of plays (throughout the period) resulted in the following plays/scenes providing the basis for sets of questions in the two versions:

1996

- Julius Caesar, Act Two, Scene One, Lines 1-228.
- Julius Caesar, Act Three, Scene Two, Lines 1-221.
- A Midsummer Night's Dream, Act Four, Scene One, Lines 43-211.
- A Midsummer Night's Dream, Act Five, Scene One, Lines 106-348.
- Romeo and Juliet, Act Three, Scene Three, the complete scene.
- Romeo and Juliet, Act Four, Scenes One to Four, the complete scenes.

[1] In both versions pupils failing to achieve L4 in the test would be awarded L3, if they gained sufficient marks. In 2001, pupils with a teacher assessment of level 3 or below were advised not to take the test and only their TA level was reported. Of those taking the 2001 test, any not reaching the L3 cut-score were recorded as 'N'. In 1996 pupils could enter the tests **and** for classroom based tasks (L1 – 3) at the teacher's discretion, achieving L3 by either route - an alternative not available in 2001. In practice only a small number of pupils would be affected. As a further consequence, level related assessment criteria in the mark scheme were changed to exclude references to level 3.

2001

- Henry V, Act One, Scene Two, Lines 96-310.
- Henry V, Act Two, Scene Two, the complete scene.
- Twelfth Night, Act One, Scene Five, Lines 81-266.
- Twelfth Night, Act Three, Scene Four, Lines 1-167.
- Macbeth, Act Two, Scenes One and Two, the complete scenes.
- Macbeth, Act 4, Scene One, the complete scene.

Thus none of the set plays is common between the two versions, rendering it impractical to select matched groups within classes who would be able to complete both versions, even if we could have persuaded schools in Northern Ireland to prepare[2] their children for a test based on any given play for the purpose of our experimental comparison, which would itself, quite rightly, be implausible[3].

For this insoluble logistical reason the Project was unable to undertake experimental comparisons involving Paper 2. The methodological adjustment for this will be described later.

Curricular and structural issues

Structural issues

KS3 spans the same age range (11-14) in England and NI, but NI has a largely 'selective' system post 11, in contrast with England where most children attend comprehensive schools. In the final year of KS2 in NI, transfer procedure tests are taken by those wishing to be considered for 'grammar' schools. Children who fail or do not take these attend 'secondary' schools. About 40% of NI children attend grammar schools, half of which are single sex schools. Most NI schools are also denominational, Protestant or Catholic, complicating provision still more, but such factors should not invalidate our experimental comparisons, provided that 'sampling' provides a reasonable balance of boys and girls across the ability range.

End of KS3 assessment arrangements in NI are similar to those in England, including externally set and marked formal tests in English - although NI testing arrangements for KS3 English do involve two overlapping tiers. NI children should however be familiar with testing, including tests of the sort involved here.

Curricular variations

Reading: The programmes of study for reading in England and NI generally cover the same areas and expected outcomes. Overall the level descriptions cover the same criteria, although in some cases they appear at different

[2] This same inherent problem makes the KS3 English test developer's job unusually difficult. Finding pupils as well prepared for trial test materials as they would be for their operational test is impossible, which makes it difficult to develop tests and alternative sets of questions within them of equivalent intrinsic difficulty, or to equate different tests and/or alternatives for Paper 2 in KS3 English.

[3] Especially as the study of Shakespeare is not compulsory in KS3 in Northern Ireland.

levels. Aspects of reading from the curriculum in England which do not feature specifically in the NI curriculum include:

- how and why texts have been influential and significant e.g. the influence of Greek myths, the Authorised Version of the Bible, the Arthurian Legends.
- distinguish between fact and opinion, bias and objectivity.
- range of literature:
  2 plays by Shakespeare.
  works of fiction by 2 major writers published before 1914 (from list provided).
  2 works of fiction by major writers published after 1914.
  poetry by 4 major poets published before 1914 (from list provided).
  poetry by 4 major poets published after 1914.

Writing: Writing covers broadly the same areas and expected outcomes in NI and England. Overall the level descriptions cover the same criteria, although in some cases they appear at different levels. Aspects of writing from the curriculum in England which do not feature specifically in the NI curriculum include:

- use of rhetorical devices.
- structure of phrases and clauses and how they can be combined to make complex sentences e.g. co-ordination and subordination.
- cohesion of texts, openings and conclusions in different types of writing e.g. through the use of verb tenses, reference chains.
- use of appropriate grammatical terminology to reflect on the meaning and clarity of individual sentences e.g. nouns, verbs, prepositions, conjunctions, articles.
- autobiography.
- prospectuses, minutes.
- editorials, campaign literature.

Handwriting: 'Joined writing' is specified in the level descriptions at levels 3, 4 and 5 in England. Although this is not explicit in the NI curriculum the mark schemes for the tests give credit for 'fluent and legible' handwriting, so any actual curricular variation will not invalidate comparisons. Level 5 in England includes the ability to adapt handwriting to a range of tasks where appropriate, which does not appear in the NI curriculum documentation.

Spelling: The curriculum in England is more specific including

- increase knowledge of roots of words and derivations, including stem, prefix, suffix, inflexion.
- apply knowledge of word formation.

Although not specifically mentioned, these could however be seen as subsumed in general terms within the NI curriculum documents.

Analysis of the content of Paper 1 in both the 1996 and 2001 versions by project staff suggested that there are no issues arising from the curricular variations discussed above which would invalidate our experimental comparisons. The reactions by teachers from the schools participating in the Project to the test materials (see below) should shed further light on this issue.

Teachers' opinions

Nine of the ten schools administering the KS3 English tests returned the Project's questionnaire seeking their reactions. Of these nine schools, six reported no problems with the administration of the tests, with another school commenting that the administration instructions were clear and easy to follow.

At a general level, the great majority of schools were content with the match between the 1996 version of the QCA test and the NI curriculum, eight out of nine recording that it matched 'very' or 'quite' well and only one response suggesting there was a poor match. In contrast, although the majority did feel that the content of the 2001 version was a reasonable reflection of the NI curriculum, four schools felt that there was a poor match.

When asked to describe any problems they felt their students might have with the 1996 test, four teachers commented that 'advertisements' as stimuli (i.e. Q3 max 11 marks) would normally[4] be used at KS4 in Northern Ireland and therefore, this task may prove demanding for KS3 pupils. This curricular variation had not been picked up by the Project's desk analysis of curriculum documents.

The 2001 version was criticised for being too long[5] and it was asserted that pupils had struggled with the time allocation. Two schools were also critical of the 2001 paper for being too narrowly focussed on 'writer's craft skills' and hence, in their opinion, the test did not reflect the breadth of study of English in Northern Ireland. But neither of these criticisms stemmed from curricular disparities, instead reflecting more general issues which would affect pupils in England equally, and hence they should not affect the validity of our experimental comparisons. It would appear that the schools' apparent suggestion that the 2001 version matched the NI curriculum less well than the 1996 version stems from such general issues, rather than differences in what should be taught and learned.

Similarly, teachers' comments on both versions raised the issue of 'unfamiliarity' caused by some differences between the structure of the QCA tests and the operational end of KS3 tests in NI. For example, NI pupils would be more familiar with a low-mark, 'warm-up' question as the opening question and would also only be required to complete two sections, rather than three. Another school commented that NI KS3 English tests target level 3 and they therefore excluded many lower ability pupils from the experimental tests. Other comments were about features of specific questions which were liked or disliked. But such issues and comments were not pertinent to whether or not the 1996 and 2001 versions of KS3 English could be seen as equally 'fair' for Northern Ireland children, our current concern.

Overall, it would seem reasonable to conclude that they were equally fair, subject to the large proviso that this applied only to Paper 1, the only

[4] But note that advertisements are explicitly mentioned in NI KS2 English documentation and persuasive writing features in the NI Programmes of Study for both KS2 and KS3.
[5] Although in fact the stimuli and questions in the two versions are quite similar in this respect.

component where experimental comparisons were feasible in practice. The Project's desk analyses of the tests and curriculum documentation suggested that NI pupils should be able to attempt both the 1996 and 2001 versions without undue difficulties arising and whilst staff in some of the schools involved appear less enamoured of aspects of the 2001 version of Paper 1, their reasons for disliking it are not rooted in the curriculum and would not indicate a bias in our comparisons.

The Data

Recruitment of schools to assist in the Project took account of the diversity of secondary schools in NI, looking for a balanced mix (rather than a random sample) of schools from throughout Northern Ireland[6], including grammar and secondary schools etc. Ten schools agreed to assist the Project and administered Paper 1 of the 1996 and 2001 versions of the QCA KS3 English test. These included six mixed schools, two boys' schools and two girls' schools.

Completed test papers were returned to Cambridge for marking, together with the pupils' operational NI KS3 English test results. Four experienced KS3 English test markers marked the scripts, after being briefed by the Operational Lead Marker for England (whose experience dated back to 1996) - using the operational co-ordination scripts for both versions. Marking was co-ordinated by a team leader who also had experience of operational marking in 1996 as well as 2001, and each marker was given a random apportionment of scripts from each version.

After data entry and cleaning, statistical analyses investigated variations in the marks awarded by the different markers. Perhaps unsurprisingly given the nature of the subject and the judgemental basis for marking, this detected statistically significant differences between the marks awarded by different examiners. Fortunately however there was no significant interaction between differences between markers and the two versions of the test, so that whatever marker effects there are may be taken as applying equally to the 1996 and 2001 versions and hence not invalidate the comparisons we wish to make. For this reason no scaling was applied to the relative severity/lenience between markers.

Further data checks considered the numbers, means and variances of the groups assigned to each version of Paper 1 within each school. These were consistent with the instructions to schools designed to allocate random groups by spiral allocation within gender.

Some children were absent when one of the test papers was administered and schools were unable to provide details of performance by all pupils on their NI end of KS3 English test, but full data were obtained for a total of 1,026 pupils. Table 2.7.2 shows the proportions of girls and boys, which are consistent between versions.

[6] Initial contacts with schools were made by NICCEA, on the behalf of the Project.

Table 2.7.2 Numbers taking each version of KS3 English by gender

	female	*male*	*total*
1996	282 (53.8%)	242 (46.2%)	524 (100%)
2001	273 (54.4%)	229 (45.6%)	502 (100%)
n 1026	*Chi-square Likelihood Ratio 0.33, df 1, n.s.*		

How well matched were the groups assigned to the 1996 and 2001 versions of the QCA KS3 English tests? The distributions of NI KS3 English test levels for the groups assigned to the 1996 and 2001 versions are shown in table 2.7.3. They are very similar, as would be expected from randomly assigned groups, and can be regarded as well matched with respect to achievement in English on their operational NI KS3 English tests.

Table 2.7.3 NI KS3 English test levels (by version of QCA KS3 Sc taken)

	L2	*L3*	*L4*	*L5*	*L6*	*L7*	*L8*		*total*
1996	0.2%	1.5%	12.4%	24.2%	38.0%	21.0%	2.7%	100%	n 524
2001	0.0%	0.8%	12.7%	23.9%	38.0%	22.1%	2.4%	100%	n 502
n 1026	*Chi-square Likelihood Ratio 2.80, df 6, n.s.*								

How do the results on the 1996 and 2001 versions of the test compare?

Can we compare Paper 1 marks from the two versions?
Mean marks obtained by the pupils assigned to the 1996 and 2001 forms of KS3 English Paper 1 are shown in table 2.7.4.

Table 2.7.4 Paper 1 mean marks by version & gender

Gender	*Version*	*mean*	*sd*	*n*
boys	*1996*	26.07	9.95	242
	2001	25.22	9.67	229
girls	*1996*	36.87	10.69	282
	2001	33.53	10.56	273
all	*1996*	31.88	11.67	524
	2001	29.74	10.97	502

n 1026
ANCOVA analysis
NI KS3 English F = 699.02, df 1, significance <0.001
Gender F = 215.03, df 1, significance <0.001
Version F = 21.65, df 1, significance <0.001
Interaction of Gender and Version F = 3.02, df 1, n.s.

Girls obtained much higher marks than boys, no matter which version of Paper 1 they were assigned to, but both boys and girls assigned to the 1996 version of the test were awarded higher marks, on average, than those assigned to the 2001 version. The difference (of 2.14 marks overall) between versions was statistically significant, but what might this mean? Whilst it might be tempting to infer that the lower marks obtained via the 2001 version suggest that this version was the more severe it would be wrong to do so. In fact, in itself it tells us nothing at all.

Yes, those taking the 2001 version obtained lower marks, and, yes, the grade related mark schemes for the 1996 and 2001 versions were in essence very similar. But the marks pupils will have been awarded will inevitably also be

influenced by the nature of the stimulus materials and questions set in the two versions. These might, and indeed probably will, vary in inherent difficulty; making it easier or harder for children's answers to display the qualities which the mark schemes reward. Such variation is indeed more likely here than in tests in other subjects, because of the way in which sub-sets of KS3 English questions are set in association with extended text-based stimuli. The texts themselves will vary in complexity and the range of questions which can be set are inherent in the nature and detail of each stimulus. In such circumstances it is much more difficult for test developers to control test difficulty than when each mark is awarded via a question which is independent of all others. It is precisely because different versions will vary in accessibility that QCA must set different threshold marks for the award of the various levels each year and it is the combination of test difficulty and thresholds which govern test standards.

Moreover, comparisons cannot be made by simply applying the thresholds from 1996 and 2001 and comparing the resulting levels, because KS3 English thresholds relate only to the total marks obtained from the combination of Paper 1 and Paper 2. They were only set at this level of aggregation. Separate thresholds for Papers 1 and 2 just do not exist. And we have already detailed the curricular and logistical reasons why it is not feasible, in Northern Ireland or England or anywhere else, to find matched groups of pupils equipped to take different versions of Paper 2 of QCA's KS3 English tests. So how can we compare the two versions?

Estimating Paper 2 marks in order to compare the two versions
Our methodology for experimental comparisons requires a method of estimating Paper 2 marks for the groups assigned to the 1996 and 2001 versions of Paper 1 in our experiment. Estimated Paper 2 marks can then be combined with actual Paper 1 marks, test levels can be derived from these and comparisons made.

To estimate Paper 2 marks we made use of statistical information about the relationship between marks on Paper 1 and Paper 2 awarded to pupils from large samples of children from schools in England who had taken part in the final trials of the 1996 and 2001 tests respectively, conducted by QCA's KS3 English test development agency - who supplied the necessary data.

Regression analyses of the data from the 1996 and 2001 versions' final trials fitted linear models which best predicted the average Paper 2 mark obtained by pupils with any given mark on Paper 1. These regression analyses were conducted for boys and girls separately, given the gender differences in achievement in English. The regression equations obtained are shown below.

1996 version (Final Trial n 1262)
- boys (n 579) Paper 2 = 5.393 + 0.349 Paper 1
- girls (n 682) Paper 2 = 7.337 + 0.339 Paper 1

2001 version (Final Trial n 521)
- boys (n 245) Paper 2 = 4.796 + 0.470 Paper 1
- girls (n 276) Paper 2 = 3.601 + 0.559 Paper 1

'Predicted' test levels for each pupil in our experimental comparisons were consequently derived by, first, using the above regressions of Paper 2 scores on Paper 1 scores, to 'predict' a Paper 2 score for each participant in our experiment.

These '*Predicted* Paper 2 totals' were then added to pupils' Paper 1 scores to produce '*Predicted* Total Marks' for the two versions of KS3 English. The operational threshold marks from 1996 and 2001 were then applied to these predicted total marks to generate a '*Predicted* KS3 English Level' for each pupil, enabling subsequent comparisons between those for pupils assigned to the two versions of Paper 1.

This methodology is far from ideal. There is for instance no empirical basis for the implicit assumption that the relationships between Paper 1 and Paper 2 marks in the final trials would be replicated in other groups. We would acknowledge that this approach is speculative, but it was the only one at our disposal.

Comparing 'predicted' test levels

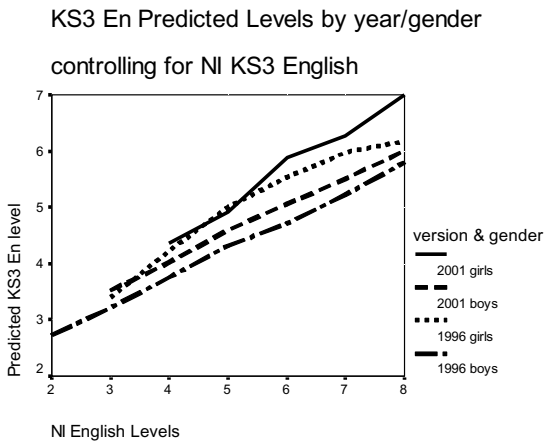
The distributions of predicted test levels associated with the two versions of the QCA KS3 English test are shown in table 2.7.5.

Table 2.7.5 Predicted QCA KS3 English test levels (by version of QCA KS3 En taken)

	L2	*L3*	*L4*	*L5*	*L6*	*L7*		*total*
1996	0.2%	6.5%	23.3%	37.6%	27.5%	5.0%	100%	n 524
2001	0.8%	2.4%	19.5%	35.1%	29.9%	12.4%	100%	n 502
n 1026	*Chi-square Likelihood Ratio 31.52, df 5, significance <0.001*							

The distribution of predicted test levels for those assigned to the 2001 version (where 42.3% reach level 6 or above) is clearly different to that for pupils assigned to the 1996 version, where only 32.5% do so.

Figure 2.7.1 Regression of Predicted KS3 English Levels on NI KS3 English



To portray these data graphically, Figure 2.7.1 shows the linear local regressions of the predicted levels for boys and girls assigned to each version

on the levels they achieved in their operational NI KS3 English tests. Girls at any given level of achievement in their NI tests in English on average obtained higher predicted levels in the QCA test than boys of equivalent ability, irrespective of the version they were assigned to. But the more able girls (i.e. those with NI KS3 English level 6 or above) assigned to the 2001 version tended to receive higher predicted levels than those of equivalent ability assigned to the 1996 version. Likewise boys (in their case throughout the ability range) assigned to the 2001 version obtained, on average, better predicted results than those assigned to the 1996 version.

Table 2.7.6 shows the table of means and standard deviations of predicted test levels associated with figure 2.7.1. Also shown is an Analysis of Covariance (ANCOVA) evaluating the statistical significance of differences in the predicted test levels awarded to those assigned to the 1996 and 2001 versions of Paper 1, whilst controlling for gender and for any effects relating to differences in the ability in English (estimated via NI KS3 English test levels) between the groups assigned to the two versions (although we have already established that the latter effects must be small).

Table 2.7.6 Mean & sd of predicted QCA KS3 En test levels, by test version and gender

		mean level	*sd*	*n*
1996	*girls*	5.43	0.89	282
	boys	4.51	0.87	242
	all	5.01	0.99	524
2001	*girls*	5.65	1.00	273
	boys	4.83	0.90	229
	all	5.28	1.04	502
Total		5.14	1.02	1026

ANCOVA analysis
NI KS3 English F = 593.45, df 1, significance <0.001
Gender F = 217.82, df 1, significance <0.001
Version F = 30.83, df 1, significance <0.001
Interaction of Gender and Version F = 0.22, df 1, n.s.

This confirms that girls tended to achieve better results in English than boys and indicates that the difference in predicted levels between pupils assigned to the 1996 and 2001 versions (amounting, on average, to 0.27 of a level) in this experimental comparison was statistically significant.

The effects of question choice in Writing

A further issue, which might complicate comparison of test standards in different versions, concerns the equivalence of the options available within Question 4 of Paper 1 in each version: i.e. the writing tasks. Is it equally easy for pupils to gain marks on each of the options available within each version of the test?

Table 2.7.7 shows the average marks obtained by boys and girls in our experimental comparisons choosing each of the writing tasks available in the 1996 and 2001 versions of Paper 1, together with results from ANCOVA evaluating the significance of differences in marks between options, whilst controlling for gender and ability in English (the latter estimated via NI KS3 English test levels).

Table 2.7.7 Writing Options

	n	*mean*	*sd*	*female (mean / n)*	*male (mean / n)*
1996					
4a A place under threat (persuasive)	193	14.95	7.4	17.86 / 100	11.83 / 93
4b Someone frightened (fiction)	149	16.85	7.5	19.22 / 96	12.57 / 53
4c Dangerous sports etc. (argument)	182	15.15	6.9	18.14 / 86	12.48 / 96
2001					
4a An unforgettable experience (fiction)	210	14.79	7.1	17.56 / 110	11.74 / 100
4b A natural disaster (report)	218	14.94	6.5	16.90 / 120	12.55 / 98
4c Sales leaflet (persuasive)	74	15.00	6.7	16.05 / 43	13.55 / 31

ANCOVA analyses

1996 NI KS3 English F = 281.85, df 1, significance <0.001
n 524 Gender F = 89.20, df 1, significance <0.001
Option F = 0.31, df 2, n.s.
Interaction of Gender and Option F = 0.31, df 2, n.s.

2001 NI KS3 English F = 333.99, df 1, significance <0.001
n 502 Gender F = 48.37, df 1, significance <0.001
Option F = 0.83, df 2, n.s.
Interaction of Gender and Option F = 1.18, df 2, n.s.

The writing options within the 1996 version were about equally popular, although girls were rather more likely to choose question 4b, a fiction writing assignment, whilst boys showed some preference for 4c, which involved argumentative writing. The apparent differences in mean marks achieved by the groups selecting these different options are in fact illusory, being accounted for by variations in the level of ability of the pupils concerned. Differences in mean marks between 1996 writing options were hence not statistically significant.

Those assigned to the 2001 version were much more likely to select questions 4a (fictional writing) or 4b (a report) than 4c (persuasive writing), but mean marks on the different options are fairly similar and were again not statistically significant.

Thus (as proved the case for similar analyses of question choice in writing assignments in KS2 English) those responsible for setting and marking these alternative writing options were successful in the not insignificant challenge of managing this without unfairness to groups of children selecting particular questions.

Do standards in 1996 and 2001 differ?

Our analysis of differences in predicted levels achieved via the 1996 and 2000 versions of KS3 English concluded that those assigned to the 2001 version might be likely, on average, to receive a predicted level 0.27 higher than those assigned to the 1996 version.

But before reflecting on this, we wish to present the results of an allied small-scale study, carried out to investigate a concern initially expressed by the KS3 English test development agency, before the Project's experimental comparisons took place. Their concern led us to 'build in' a small-scale additional experiment to test an hypothesis implicit within it.

Might markers' expectations have changed since 1996?
The concern expressed stemmed from the perceived success of curricular changes in English induced in schools in England in recent years via the national curriculum, and bolstered by national tests. Arguably, in the early years of KS3 national testing (1996 being a prime example) the pupils concerned may have been relatively less well prepared for the range of tasks set and less likely to produce the responses sought by the KS3 English mark scheme than are today's pupils. For much of their education their teachers will not have been aware of these new targets, and even in the run up to end of KS3 tests their teachers will still have been coming to terms with what was now expected and how best to inculcate it. Pupils taking statutory tests in 2001 will have reaped the benefits from their teachers' sharp learning curve. If this is true we might expect the quality of children's work, taken as a whole, to be better and to be rewarded by higher levels, as appears to have been the case. But the concern was that in 1996, when teachers and pupils were faced by novel challenges, markers might have been more willing (than their counterparts in more recent years) to accept marginal demonstrations of some of the skills sought as sufficient to merit the marks. Arguably, now, teachers are familiar with what is required and will have worked with pupils to develop and present their skills in the testing environment, so that they would expect to see them clearly evident in children's work. The same might then hold for KS3 English test markers, most of whom are also teachers.

Is it possible that by 2001, such raised expectations might mean that markers would expect clearer evidence before awarding some marks than their counterparts in 1996 would have done?

Our briefings of the markers for this experiment used the mark schemes and exemplar 'co-ordination' scripts used operationally in both years, but even this might not have been sufficient to insulate them from 'expectation creep'.

A small-scale re-marking study
QCA supplied us with copies of all the marked 1996 Paper 1 scripts available in their archives. There were only 24 such scripts (12 boys and 12 girls), less than would really be desirable, but fortunately they were drawn from throughout the whole mark range.

Immediately after our (n = 4) markers had finished marking the scripts (for both the 1996 and 2001 versions) for our experimental comparisons, we supplied each of them with photocopies of the 24 archive scripts from 1996, which had been 'cleaned' to remove all the original marks and annotations. They were simply asked to treat them exactly as they had the experimental 1996 version Paper 1 scripts they had just completed. We were then able to compare their marks with those originally awarded operationally, in 1996.

Table 2.7.8 Re-marking study: means, standard deviations & correlations (n 24)

	mean	*sd*	$r_{x,96}$	$r_{x,re1}$	$r_{x,re2}$	$r_{x,re3}$	$r_{x,re4}$
*1996 live marking***	28.25	14.8					
re-marker 1	23.75	10.8	0.91*				
re-marker 2	22.50	12.4	0.92*	0.92*			
re-marker 3	25.58	12.1	0.94*	0.93*	0.94*		
re-marker 4	24.96	11.0	0.94*	0.94*	0.93*	0.98*	
average of re-marks	24.20	11.3	0.95*	0.97*	0.97*	0.99*	0.98*

* correlation significant at <0.001

** The mean 1996 live mark was significantly different (at <0.001) from the mean for every re-marker in this study.

Table 2.7.8 presents the means and standard deviations of the operational 1996 marks for the 24 Paper 1 archive scripts, together with those for the four (re)markers involved and their average. It also shows the correlations between each of these sets of marks.

The correlations observed are all greater than 0.9 and hence highly significant. Correlations of between 0.91 and 0.94 are observed between the live marks and the four re-markings, whilst the correlations amongst the markers in this study range from 0.92 to 0.98. Given the judgemental nature of marking in this subject, these correlations would normally[7] be regarded as very re-assuring estimates of mark/re-mark reliability, both for the quality of operational marking and for this experiment.

The means of the marks awarded by all four markers from this experiment were however lower than the 1996 live marks. The magnitude of the differences observed were statistically significant in every case and, on average, the marks awarded by our experimental markers were 4.05 lower than those awarded operationally in 1996: a very substantial difference.

This would seem to confirm the implicit hypothesis within the argument that teachers' and markers' expectations have risen since 1996. If correct, this view would imply that today's markers would be likely to award lower marks than those marking operationally in 1996, despite the apparent similarity of the mark schemes used across this period - just what we have observed.

[7] But we should avoid over complacency. The high correlations show that the four markers ranked the 24 scripts in very similar orders, but the means of the Paper 1 total marks awarded ranged from 22.50 to 25.58, a difference of 3.08 marks. Despite the fact that very few scripts were involved, of the six paired comparisons possible between the four markers (1 with 2,3 & 4; 2 with 3 & 4; 3 with 4), statistical analyses suggested that the differences in mean marks awarded were statistically significant at the 0.01 level in two instances (marker 2 v marker 3 and marker 2 v marker 4).

Conclusions

The conclusions we might reach about relative standards for the 1996 and 2001 versions of KS3 English depend upon the credence we give to the re-marking study reported above as well as to the experimental comparisons described earlier. The re-marking study involves only 24 archived scripts from 1996, but these were all that existed and the differences in marks awarded here and in 1996 were large: enough to be highly significant statistically. The argument and the evidence seem plausible, despite the small scale.

So if we accept the validity of the argument and test of the expectation creep hypothesis we are, in effect, suggesting that the marks awarded to those assigned to the 1996 version in our experimental comparisons were lower than those which would have been awarded operationally by 1996's markers. Our best estimate of the extent of any under-marking involved is the difference between the operational marks for the archive scripts and the average of the four re-markings: i.e. about four marks.

How much is four marks on Paper 1 in terms of its effect on test levels? Imagine that we were to have given everyone assigned to the 1996 version a four mark bonus - what difference might it have made? In 1996 the threshold for level 3 was 14 marks, whilst that for level 6 was 74 marks, so levels 3 to 6 spanned a total of 60 marks - an average of 15 marks per level. On this scale 4 marks represents 0.27 of a level, which thus represents our estimate of the potential severity of our experimental marking of the 1996 version due to expectation creep since 1996.

The experimental evidence itself could also be challenged on methodological grounds because of our inability to collect experimental data regarding Paper 2 and the consequent need to predict Paper 2 marks from Paper 1 and use these to derive, and compare, predicted test levels. But in any event the 0.27 of a level disparity in marking arising from expectation creep exactly cancels out the evidence from our experimental comparison, that those assigned to the 2001 version obtained levels which were, on average, 0.27 higher than those assigned to the 1996 version.

It would seem safest to conclude that the balance of the evidence is that this investigation provides no indication that the standards arising from the combined effects of the question papers, mark schemes and level thresholds of the 1996 and 2001 versions of KS3 English tests are any different.

2.8 KS3 Mathematics experimental comparison: 1996 v 2000

Historical Trends

The historical pattern of KS3 Mathematics results is provided in table 2.8.1 and this shows how they have improved in the last few years. In 1996 33% of the cohort obtained Level 6 or better but by 2000 42% reached such levels. The pattern is not one of steady progress. The 1997 results represented a substantial step forward but in 1998, when the mental arithmetic element was first introduced to KS3 Mathematics, results lost ground, subsequently regained in 1999. Further progress followed in 2000, especially at Levels 7 and 8.

Table 2.8.1 KS3 Mathematics test results 1996-2000 (% of cohort at each level*)

	below 3	*Level 3*	*Level 4*	*Level 5*	*Level 6*	*Level 7*	*Level 8*	*EP*
1996	3%	11%	23%	23%	22%	10%	1%	0%
1997	2%	10%	22%	23%	25%	11%	1%	0%
1998	2%	11%	22%	24%	23%	11%	2%	0%
1999	3%	9%	21%	24%	24%	12%	2%	0%
2000	2%	9%	20%	24%	23%	16%	3%	0%

* rows may not total 100% because those absent or disapplied are excluded

The validity of experimental comparisons in schools in Northern Ireland

Variations in style and content between the 1996 and 2000 versions

The 1996 and 2000 versions of the test both had the same four overlapping ability related tiers (T3-5, T4-6, T5-7 & T6-8) and, in both versions, each tier had two essentially parallel 1 hour written papers, with the style and content of questions in these also remaining very similar.

But there was one major difference between the 1996 and 2000 versions. An additional element, a Mental Arithmetic test, was included in the 2000 version. There were three versions of this. A and B were equivalent forms targeted at levels 4-7, so that schools could use either or both. Form C was an easier test for those taking T3-5 in the written papers. All three consisted of a 20 minute audio-tape containing 30 questions. The Mental Arithmetic marks were added to the marks gained on the written papers (and carried 20% of the total marks available).

Curricular and structural issues

Structural issues
KS3 spans the same age range (11-14) in England and NI, but NI has a largely 'selective' system post-11, in contrast with England where most children attend comprehensive schools. In the final year of KS2 in NI, transfer procedure tests are taken by those wishing to be considered for 'grammar' schools. Children who fail or do not take these attend 'secondary' schools. About 40% of NI children attend grammar schools, half of which are single sex schools. Most NI schools are also denominational, Protestant or Catholic, complicating provision still more, but such factors should not invalidate our

experimental comparisons, provided that 'sampling' provides a reasonable balance of boys and girls across the ability range.

End of KS3 assessment arrangements in NI are similar to those in England and include external tests in Mathematics involving ability related tiers. However the NI tests do not include a mental arithmetic element, so this would be unfamiliar to pupils involved in the experimental testing, which might depress results on the 2000 version of the test.

Curricular issues
Curriculum documentation was reviewed by project staff to identify topics mentioned explicitly in QCA's documents but not found in NICCEA's, which might then bias experimental comparisons if such topics were more prevalent in one version than the other. Those identified were:

Number & algebra
- multiply two linear expressions
- solve inequalities in two variables

Shape, space & measures
- solve bearing problems
- determine locus of an object according to a given rule, including, where appropriate, using practical methods and devising instructions for a computer to produce desired shape/path

Handling data
- draw inferences from statistics
- take account of bias
- evaluate results critically and develop an understanding of the reliability of results
- recognise that inferences from analyses may suggest further questions to investigate
- engage in practical and experimental work in order to appreciate some of the principles governing random events
- select and calculate or estimate appropriate measures of spread, including the interquartile range applied to discrete, grouped and continuous data
- understand how to calculate the compound probability of two independent events, given their own probabilities

These are topics where NI pupils might be less well prepared than English pupils, if NI teachers do not anyway recognise their importance and include them in their teaching. Some look especially likely to be favoured by teachers, notably some listed under 'handling data'.

The 1996 and 2000 question papers were also carefully examined to try to determine how many marks might be affected by these curricular variations. Few questions seem to address these areas.

In the 1996 version one question (referring to the compound probability of events and carrying 4 marks) appears problematic. Two others (involving 8 marks) used bearings as a context, but these could also be answered by utilising trigonometry.

In the 2000 version only one question (again testing the compound probability of events and carrying 4 marks) was affected.

With so few questions affected and an equal balance between versions there seems no reason why experimental comparisons in NI schools should be invalidated by the curricular differences identified.

Teachers' opinions
Staff from schools taking part were asked to complete questionnaires seeking their views on the fairness of the two versions of the test for their pupils, generally and in relation to particular questions. Ten of the fourteen participating schools returned questionnaires. In some cases mathematics staff had completed a 'composite' questionnaire but other schools returned more than one, so that in all nineteen questionnaires were returned.

Replies suggested that schools had coped effectively with the especially complex test administration involved in this comparison. Five commented on this issue and we must thank all the teachers involved for their time and efforts.

Teachers were asked how well the 1996 and 2000 versions of KS3 Mathematics reflected the content of the NI Mathematics curriculum. Table 2.8.2 summarises their replies, which suggest most thought them a reasonable match to their pupils' experience.

Table 2.8.2 How well do the 1996 and 2000 tests reflect the NI curriculum?

	1996	*2000*
very closely	6	6
similar	10	11
not at all	2	0
no response	1	2

n 19

Only six questions from the 1996 version of the test attracted comments suggesting that NI pupils would be inadequately prepared because of curricular differences between NI and England and only two of these attracted comments from more than one respondent.

Seven of the questions in the 2000 version had comments mentioning curricular differences and only one of these attracted the attention of more than one respondent.

Other comments (relating to 13 questions on the 1996 test and 10 questions on the 2000 version, with some overlap between these and the above questions attracting comments with a curricular element) often suggested questions were likely to prove challenging, or referred to specific features. One widely held general view (expressed by 11 respondents) concerned the style of the tests, which were seen as 'wordy' by contrast with their NICCEA equivalents, but this applied to both versions.

The overall impression was that the teachers were of the view that both versions of the test were reasonably appropriate for their pupils. The few curricular concerns raised were evenly distributed between the 1996 and

2000 versions, and so seemed unlikely to invalidate comparisons between them.

The Data

Sixteen NI secondary schools (selected to include a balance of grammar and secondary schools and of boys', girls' and mixed schools, from throughout Northern Ireland) were approached by NICCEA on the Project's behalf and asked to participate. Fourteen agreed to do so and administered the 1996 and 2000 versions of the KS3 Mathematics test to their Secondary 3 cohorts in June 2000, shortly after the completion of operational testing (in both England and NI). In all 1,490 pupils were involved.

Test materials and instructions for test administration were supplied to the schools. Instructions were based on QCA's operational versions, edited by project staff to facilitate the simultaneous administration of two versions of the test and the other administrative complications the experimental design incurred.

Test administration was in this instance - unavoidably - complex. Schools needed to allocate pupils to the tier[1] most appropriate to their abilities and distribute test papers accordingly. This involved the two written papers for each pupil (where tiers 3-5, 4-6, 5-7 and 6-8 were available) and the Mental Arithmetic tests (where T3-5 pupils took version C and all others took version A). Note that because the 1996 test did not include a mental arithmetic element all those taking part took the appropriate 2000 Mental Arithmetic test. In addition, schools were asked to use spiral allocation within gender to allocate pupils to either the 1996 or 2000 versions. Ensuring that the correct test papers were in the right hands every time was not simple!

Following testing, schools returned papers to Cambridge for marking, together with details of gender and operational NI KS3 Mathematics test results. A team of 3 experienced markers was briefed by a senior member of the operational KS3 marking team with experience of both 1996 and 2000 marking. Each marker was allocated a stratified random sample of scripts from each school, including both the 1996 and 2000 versions and all test components and tiers. Statistical quality control comparisons confirmed that there were no significant differences between the marks awarded by the three markers. Data cleaning involved range and total checks, and checks on the numbers of boys and girls allocated to each version within each school and the mean and spread of their marks. These suggested that two schools may not have followed the instructions for spiral allocation of random groups to versions correctly and their data were therefore excluded from subsequent analyses. It was also necessary to exclude data from children who were absent from one or other of the papers, together with a small number who had taken the 'wrong' version of a paper due to administrative errors.

[1] Schools were at least used to this problem, NI tier arrangements being similar to QCA's.

Usable data were obtained for a total of 1,069 pupils from 12 schools. These included 5 secondary schools (1 girls only, 1 boys only and 3 mixed) and 7 grammar schools (1 girls, 2 boys and 2 mixed). Most schools divided their children between the two or three tiers appropriate to their intake's ability range, but two grammar schools entered all their pupils for a single tier - in one case, a girls' school, T5-7; in the other, a boys' school, T6-8. This inevitably produced some gender imbalances within these tiers.

Were the groups assigned to the two versions equivalent?

Given the complex structure of KS3 Mathematics tests, answers to this question must take account of tiering arrangements. Gender effects are also relevant. Usable data were available for 562 girls (283 taking the 1996 version and 279 the 2000 version) and 507 boys (265 taking 1996 and 242 taking 2000). Table 2.8.3 summarises the numbers of boys and girls in each tier by test version, whilst table 2.8.4 shows their NI KS3 Mathematics levels.

Table 2.8.3 Gender by test version within tiers

tier	*version*	*girls*	*boys*		*n*
3-5	*1996*	47.0%	53.0%	100%	115
	2000	44.1%	55.9%	100%	102
	total	45.6%	54.4%	100%	217
4-6	*1996*	47.0%	53.0%	100%	117
	2000	48.5%	51.5%	100%	99
	total	47.7%	52.3%	100%	216
5-7	*1996*	62.4%	37.6%	100%	202
	2000	68.6%	31.4%	100%	191
	total	65.4%	34.6%	100%	393
6-8	*1996*	42.1%	57.9%	100%	114
	2000	42.6%	57.4%	100%	129
	total	42.4%	57.6%	100%	243

T3-5 Chi-Square Likelihood Ratio 0.18, df 1, n.s.
T4-6 Chi-Square Likelihood Ratio 0.05, df 1, n.s.
T5-7 Chi-Square Likelihood Ratio 1.68, df 1, n.s.
T6-8 Chi-Square Likelihood Ratio 0.01, df 1, n.s.

Boys formed a slightly higher proportion of those allocated to both versions in both T3-5 and T4-7, but the variations in the proportions of boys and girls assigned to the two versions were not statistically significant. In T5-7 there were markedly more girls than boys, whilst this was reversed in T6-8. This arose from the single sex schools' entry policies described above. But these gender imbalances applied to both versions of the test with differences between them again not statistically significant. Gender imbalances should not therefore invalidate comparisons between the 1996 and 2000 versions.

The NI assessments shown in table 2.8.4 provided a common yardstick to compare the mathematical ability of the groups assigned to the two versions of the test. They suggest that in all four tiers the groups appeared well matched in this respect also, as the experimental design intended, with any variations between versions not proving statistically significant.

Table 2.8.4 NI KS3 Mathematics levels by version within tiers

Tier		*NCL 2*	*NCL 3*	*NCL 4*	*NCL 5*	*NCL 6*	*NCL 7*	*NCL 8*		*n*
3-5	*1996*	0.9%	18.3%	57.4%	20.9%	1.7%	0.9%		100%	115
	2000		21.6%	53.9%	24.5%				100%	102
	total	0.5%	19.8%	55.8%	22.6%	0.9%	0.5%		100%	217
4-6	*1996*			0.9%	55.6%	43.6%			100%	117
	2000			2.0%	62.6%	33.3%	2.0%		100%	99
	total			1.4%	58.8%	38.9%	0.9%		100%	216
5-7	*1996*				1.0%	35.1%	63.4%	0.5%	100%	202
	2000			0.5%		35.1%	63.4%	1.0%	100%	191
	total			0.3%	0.5%	35.1%	63.4%	0.8%	100%	393
6-8	*1996*					9.6%	47.4%	43.0%	100%	114
	2000					7.8%	50.4%	41.9%	100%	129
	total					8.6%	49.0%	42.4%	100%	243

T3-5 Chi-Square Likelihood Ratio 5.8, df 5, n 217, n.s.
T4-6 Chi-Square Likelihood Ratio 5.6, df 3, n 216, n.s.
T5-7 Chi-Square Likelihood Ratio 4.5, df 4, n 393, n.s.
T6-8 Chi-Square Likelihood Ratio 0.4, df 2, n 243, n.s.

How do results on the 1996 and 2000 versions compare?

The distributions of KS3 Mathematics test levels achieved by the groups assigned to the 1996 and 2000 versions, for all tiers combined, are shown in table 2.8.5. Those assigned to the 2000 version obtained better results than those assigned to the 1996 version; the differences being statistically significant. Given that we have established that these groups were well matched with respect to ability in Mathematics, as estimated by NI KS3 Mathematics test results, this suggests that the equivalence of standards between these two versions is questionable.

Table 2.8.5 KS3 Mathematics levels by test version (all tiers)

	NCL 0	*NCL 2*	*NCL 3*	*NCL 4*	*NCL 5*	*NCL 6*	*NCL 7*	*NCL 8*		*n*
1996	0.5%	1.1%	8.8%	11.1%	14.2%	31.6%	26.6%	6.0%	100%	548
2000	0.2%		3.5%	11.9%	13.2%	26.7%	36.5%	8.1%	100%	521

Chi-Square Likelihood Ratio 34.0, df 7, n 1,069, significance <0.001

Table 2.8.6 investigates the equivalence of standards between versions within tiers, via distributions like those above.

Table 2.8.6 KS3 Mathematics levels by version within tiers

Tier		*NCL 0*	*NCL 2*	*NCL 3*	*NCL 4*	*NCL 5*	*NCL 6*	*NCL 7*	*NCL 8*		*n*
3-5	*1996*	0.9%	5.2%	41.7%	40.9%	11.3%				100%	115
	2000	1.0%		17.6%	56.9%	24.5%				100%	102
4-6	*1996*	0.9%			12.0%	44.4%	42.7%			100%	117
	2000				3.0%	44.4%	52.5%			100%	99
5-7	*1996*					5.4%	56.9%	37.6%		100%	202
	2000				0.5%		38.7%	60.7%		100%	191
6-8	*1996*	0.9%				1.8%	7.0%	61.4%	28.9%	100%	114
	2000						10.1%	57.4%	32.6%	100%	129

T3-5 Chi-Square Likelihood Ratio 26.7, df 4, n 217, significance <0.001
T4-6 Chi-Square Likelihood Ratio 8.3, df 3, n 216, significance <0.05
T5-7 Chi-Square Likelihood Ratio 33.7, df 3, n 393, significance <0.001
T6-8 Chi-Square Likelihood Ratio 5.63, df 4, n 243, n.s.

The distributions of levels achieved via the 1996 and 2000 versions reveals that (statistically) significantly better results were obtained by the groups assigned to the 2000 versions of tiers, T3-5, T4-6 and T5-7, but not by those taking the highest tier, T6-8. Similar analyses of distributions for boys and girls (not reported here) were also undertaken and such differences were evident within both gender groups.

An alternative perspective on these data can be obtained by assuming that levels (both England's and Northern Ireland's) have equal interval properties and using analyses of co-variance (ANCOVA) to assess the significance of differences in levels achieved between the groups assigned to the 1996 and 2000 versions whilst controlling for differences between genders and for the ability of the boys and girls taking the two versions of test. Table 2.8.7 provides the breakdown of mean levels by tier, version and gender, and reports the ANCOVA analyses assessing the significance of the various 'effects' within each tier.

Table 2.8.7 Mean KS3 Maths levels by tier, version and gender

Tier 3-5

		mean	*sd*	*n*
1996	*boys*	3.62	0.80	61
	girls	3.48	0.86	54
	all	3.56	0.83	115
2000	*boys*	4.11	0.84	57
	girls	3.93	0.65	45
	all	4.03	0.76	102
Total		3.78	0.83	217

ANCOVA (n 217)

NI Maths F = 69.51, df 1, significance <0.001
Gender F = 0.06, df 1, n.s.
Version F = 26.58, df 1, significance <0.001
Interaction of Gender & Version F = 0.30, df 1, n.s.

Tier 4-6

		mean	*sd*	*n*
1996	*boys*	5.26	0.92	62
	girls	5.27	0.73	55
	all	5.26	0.83	117
2000	*boys*	5.53	0.58	51
	girls	5.46	0.54	48
	all	5.49	0.56	99
Total		5.37	0.73	216

ANCOVA (n 216)

NI Maths F = 58.07, df 1, significance <0.001
Gender F = 0.09, df 1, n.s.
Version F = 9.73, df 1, significance <0.01
Interaction of Gender & Version F = 0.06, df 1, n.s.

Tier 5-7

		mean	*sd*	*n*
1996	*boys*	6.22	0.56	76
	girls	6.38	0.58	126
	all	6.32	0.57	202
2000	*boys*	6.62	0.49	60
	girls	6.59	0.54	131
	all	6.60	0.52	191
Total		6.46	0.57	393

ANCOVA (n 393) *NI Maths F = 102.34, df 1, significance <0.001*
Gender F = 5.47, df 1, significance <0.05.
Version F = 33.92, df 1, significance <0.001
Interaction of Gender & Version F = 2.41, df 1, n.s.

Tier 6-8

		mean	*sd*	*n*
1996	*boys*	6.98	1.12	66
	girls	7.31	0.51	48
	all	7.12	0.92	114
2000	*boys*	7.01	0.61	74
	girls	7.51	0.50	55
	all	7.22	0.62	129
Total		7.18	0.78	243

ANCOVA (n 243) *NI Maths F = 35.63, df 1, significance <0.001*
Gender F = 4.74, df 1, significance <0.05
Version F = 1.15, df 1, n.s.
Interaction of Gender & Version F = 0.76, df 1, n.s.

Although the girls involved in this experiment[2] achieved slightly higher mean levels than the boys overall (girls' mean level 5.77 to boys' 5.59 on the 1996 version and girls' mean 6.15 to boys' 5.92 on the 2000 version): this overall difference also reflects choice of tiers of entry, allowing patterns of gender difference within tiers to be more variable, as we can see.

In T3-5 the boys taking part did better than the girls on both versions. In both T4-6 and T5-7 gender superiority varies across test forms, but does so inconsistently. And in T6-8 the girls involved did better than boys on both versions. Such variations may in part be due to sampling effects, given the non-representative selection of schools involved, but they will also reflect differential self-selection effects in choice of entry tier, similar to those reported by Massey et al (1996). They illustrate the complexity of experimental design required for comparisons like these.

Fortunately the collection of an independent estimate of mathematics ability (NI KS3 Mathematics test results) makes ANCOVA feasible, enabling us to control for this potentially unbalancing factor. ANCOVA analyses for each tier are able to evaluate the significance of the differences between the levels

[2] Note that there is no suggestion here that these are representative samples, of either gender.

achieved by the groups assigned to the two versions of the KS3 Mathematics test, whilst controlling for both variations in the proportion of boys and girls taking the two versions and variations in their ability in Mathematics. At the same time the significance of differences between achievement by boys and girls is evaluated, as are interaction effects between gender and test version, although such matters are of secondary interest here.

In the event, gender effects observed within tiers were relatively weak, proving statistically significant only in T5-7 and T6-8, the higher ability range, where girls out-performed boys. This is as we might expect, reflecting their (or their schools') tier entry decisions and accounting for girls' overall superiority.

However the differences between the 1996 and 2000 versions of the test observed in the lower three of the four tiers proved highly significant, with those taking the 2000 version obtaining significantly higher levels, all things being equal, in T3-5 (where the difference was greatest, averaging 0.47 of a level), T4-6 (where differences averaged 0.23 of a level) and T5-7 (where differences averaged 0.28 of a level). Those taking the 2000 version also achieved higher mean levels in T6-8, but the difference observed here (on average 0.1 of a level) was within the range we might expect by chance.

The disparity in outcomes thus appears greatest for the pupils taking the lowest tier and least affects those taking the highest tier. Indeed to take this a step further, more detailed inspection of the data revealed that for the very ablest, i.e. those obtaining level 8 on their NI Mathematics tests, outcomes were similar from both the 1996 and 2000 versions of T6-8 in the QCA tests.

The effect of introducing Mental Arithmetic

The introduction of the Mental Arithmetic element to KS3 Mathematics (in 1998) illustrates the way in which changes in the curricular and/or testing regime pose a major challenge to the continuity of test standards. Such changes are a political imperative which cannot be avoided if the curricular regime is not to stagnate. In this instance a substantial new sub-domain has been introduced, perhaps drawing upon different aptitudes and certainly testing skills and knowledge which the managers of England's educational system wished to receive greater emphasis in teaching. This was a major discontinuity, but it is simply a starker illustration of the similar problems which arise from other, less obvious, adjustments in targets, as the curricular/testing regime evolves.

The changing distribution of levels achieved (a dip in 1998, followed by a recovery the next year) when Mental Arithmetic was first introduced alone suggests a disturbance to 'standards' - probably in several senses. Were teachers more effective in preparing children for Mental Arithmetic by 1999? If so test standards might have remained constant when performance standards dipped in 1998. Or were teaching and learning equally effective in both years but better results obtained in 1999 as a consequence of greater familiarity with the new element in the test? If so would we say performance standards

were the same in 1998 and 1999? And what would this imply for test standards? Think too of the potential knock-on. If teachers now spend more time on Mental Arithmetic, how can children be as good as they were at the other aspects of mathematics which previously occupied this time? So might 1998's dip in achievement have arisen partly, or wholly, as a consequence of weaker performance in the aspects of mathematics tested by the written papers? Even if we had empirical evidence across the period in question it would be very hard to disentangle such issues. Without it, it is impossible. But this shows why 'maintaining standards' is not a simple matter.

As reported above, our methodology allowed comparisons of the levels achieved by those assigned to the 1996 and 2000 versions of the test. The latter of course included a Mental Arithmetic element and the former did not. But we did also ask the children assigned to the 1996 version to take the 2000 Mental Arithmetic test, even though we did not include their performance on it in the calculation of the levels achieved, as reported in the comparisons between versions above. This 'extra' data does however enable us to explore the impact of this element in the test.

To do so we calculated a second result for each child assigned to the 1996 version: a hypothetical level 'modified' to include Mental Arithmetic. To do this:

- We first added (1996 version) pupils' marks for Mental Arithmetic to their totals from the written papers, creating 'modified totals'.
- To award levels appropriately we then required 'modified level thresholds' which also incorporated Mental Arithmetic. Fortunately, QCA's level setting methodology over the period 1998 to 2000 allowed us to calculate these. KS3 thresholds for written and Mental Arithmetic elements had been determined separately and then aggregated without any allowance for regression effects, so the calculation and application of modified cut-scores for the 1996 version, 'equivalent' to those for 2000, was straightforward. All that was necessary was to add the minimum mark required for each level in the versions of the 2000[3] Mental Arithmetic test used (A or C were taken by the children in this experiment) to the existing 1996 thresholds.
- The appropriate modified thresholds were then applied to modified total scores to calculate the modified 1996 test levels required (see table 2.8.9).

We then replicated the comparisons between versions described above, this time using the modified 1996 levels. These analyses are reported in table 2.8.8. The numbers involved in the two comparisons varied slightly because of partial absence (when the Mental Arithmetic test was taken) by some children assigned to the 1996 version, so that modified levels could not be calculated for them. There seems no reason why this should have affected the validity of the analyses.

[3] Indeed Mental Arithmetic thresholds were unchanged between 1998 and 2000, implying that it was believed that the test developers had produced Mental Arithmetic tests of equivalent difficulty across this period. The use of the same thresholds for the alternative versions (A and B) each year carries the same assumption.

Table 2.8.8 Mean modified KS3 Maths levels (inc. Mental Arithmetic for 1996 version)

Tier 3-5

		mean	*sd*	*n*
1996	*boys*	3.80	0.72	56
	girls	3.65	0.74	46
	all	3.74	0.73	102
2000	*boys*	4.11	0.84	57
	girls	3.93	0.65	45
	all	4.03	0.76	102
Total		3.88	0.76	204

ANCOVA (n 204)

NI Maths F = 68.90, df 1, significance <0.001
Gender F = 0.18, df 1, n.s.
Version F = 13.62, df 1, significance <0.001
Interaction of Gender & Version F = 0.22, df 1, n.s.

Tier 4-6

		mean	*sd*	*n*
1996	*boys*	5.50	0.67	62
	girls	5.35	0.68	52
	all	5.43	0.68	114
2000	*boys*	5.53	0.58	51
	girls	5.46	0.54	48
	all	5.49	0.56	99
Total		5.46	0.63	213

ANCOVA (n 213)

NI Maths F = 72.10, df 1, significance <0.001
Gender F = 0.58, df 1, n.s.
Version F = 2.18, df 1, n.s.
Interaction of Gender & Version F = 0.59, df 1, n.s.

Tier 5-7

		mean	*sd*	*n*
1996	*boys*	6.32	0.50	74
	girls	6.41	0.58	125
	all	6.38	0.55	199
2000	*boys*	6.62	0.49	60
	girls	6.59	0.54	131
	all	6.60	0.52	191
Total		6.48	0.55	390

ANCOVA (n 390)

NI Maths F = 97.29, df 1, significance <0.001
Gender F = 8.49, df 1, significance <0.01
Version F = 21.18, df 1, significance <0.001
Interaction of Gender & Version F = 0.72, df 1, n.s.

Tier 6-8

		mean	*sd*	*n*
1996	*boys*	6.98	1.10	61
	girls	7.26	0.49	47
	all	7.10	0.90	108
2000	*boys*	7.01	0.61	74
	girls	7.51	0.50	55
	all	7.22	0.62	129
Total		7.17	0.76	237

ANCOVA (n 237) *NI Maths F = 35.70, df 1, significance <0.001*
Gender F = 3.86, df 1, n.s.
Version F = 2.29, df 1, n.s.
Interaction of Gender & Version F = 1.40, df 1, n.s.

The modified levels for those taking the 1996 version were often higher than those obtained without the Mental Arithmetic element. In T3-5 mean modified levels were 0.18 higher than the original mean levels. In T4-6 mean modified levels were 0.17 higher and in T5-7 mean modified levels were 0.06 higher. But in T6-8 mean modified levels were fractionally lower than the original mean levels, by 0.02.

So when these modified 1996 levels were considered, differences between the groups taking the 1996 and 2000 versions were smaller than those encountered in the original comparisons. In all four tiers the groups taking the 2000 version still obtained higher levels, on average. However the difference between mean 2000 and mean 1996 levels modified to include Mental Arithmetic fell (relative to the original comparisons) in three tiers - from 0.47 of a level to 0.29 of a level in T3-5; from 0.23 to 0.06 in T4-6 and from 0.28 to 0.22 in T5-7, but rose to 0.12 from 0.10 in T6-8.

Consequently, with Mental Arithmetic included, the superiority in mean levels achieved by those taking the 2000 version of the test was less marked. Differences amongst pupils who took T4-6 were no longer statistically significant, like those amongst those taking T6-8, where the differences were small before the inclusion of Mental Arithmetic.

It would seem that the 'improvement' derived from the inclusion of Mental Arithmetic in the modified 1996 levels was strongest for those taking T3-5 and the Form (version C) of the Mental Arithmetic test targeted at this ability range. For those in this tier the inclusion of Mental Arithmetic reduced the deficit in mean levels for those taking the 1996 version (compared with those taking the 2000 version) by about a fifth of a level (for both sexes). It would therefore seem reasonable to attribute this portion of the almost half a level difference in unmodified levels between the two versions to the effects of including this new element in the test. But even the modified levels achieved by T3-5 pupils taking the 1996 version were substantially and (statistically) significantly lower than those of the group assigned to the 2000 version - after controlling for gender and ability in Mathematics.

Amongst those (all the rest) who took Form A of the Mental Arithmetic test, the effects observed were stronger in the lower tier, T4-6. Here boys gained improvements of 0.24 of a level when 1996 levels were modified by the inclusion of Mental Arithmetic. This was substantially greater than girls, whose levels improved, on average, by only 0.08 of a level. Again it seems reasonable to attribute this portion of the difference in unmodified levels between the 1996 and 2000 versions to the inclusion of Mental Arithmetic. The 'improvement' was sufficient that although mean modified levels on the 1996 version (for both boys and girls) in this tier were lower than those achieved by the group assigned to the 2000 version, the difference was not enough to be statistically significant.

In T5-7 boys taking the 1996 version gained, on average, improvements of 0.1 of a level as a result of modification to include Mental Arithmetic, whilst girls gained less - 0.03 of a level, following the same gender pattern as T4-6. But in this tier the mean levels achieved by those taking the 2000 version were so much better that even after modification of 1996 levels to include these modest gains from a Mental Arithmetic element, the difference remained statistically significant.

In T6-8 differences between the groups assigned to the two versions were comparatively small, especially for boys. Modification of 1996 levels to include Mental Arithmetic made no difference to the levels achieved by boys in this tier. Whilst the girls' (in a minority in this sample) mean levels did improve a little the differences between versions were, as originally, not statistically significant.

It may be unsafe to dwell on the gender variations noted above. Relatively small numbers are involved in these within tier comparisons and with single sex schools involved, school and gender effects may well be confounded, especially in tiers 5-7 and 6-8.

More importantly, it was the weaker pupils who benefited most from the inclusion of Mental Arithmetic in 1996 levels. What can we conclude from this? The difficulty of Mental Arithmetic tests might have changed since 1998, in some way especially affecting the weaker pupils. But this seems improbable, not least because of the empirical trials included in the test development process. It is more likely that when Mental Arithmetic was first introduced, in 1998, the thresholds set were too lenient, especially at the lower levels. This applies both to Form C (which exclusively targets the levels most affected) and Form A[4]. But although substantial, this 'mental arithmetic effect' by no means accounts for the whole of the difference in levels achieved by those assigned to the 1996 and 2000 versions.

We should also note that taking the differences between versions observed at face value may be conservative, especially in relation to Mental Arithmetic. The review of potential curricular biases noted that children in NI will not have encountered Mental Arithmetic tests in their own end of KS3 assessments

[4] Having not used Form B our experiment cannot comment on it, except to say that if it is indeed equivalent in difficulty to Form A, as the use of common thresholds implies, the conclusions reached should apply to this Form too.

and that their teachers may not have emphasised this facet of Mathematics as much as teachers in England. If so we might expect their scores on Mental Arithmetic to be depressed, at any given level of ability, relative to children from schools in England.

But if the children from NI taking part in this experiment had scored better on Mental Arithmetic, the levels achieved on the 2000 version of the test would have been higher still, increasing differences from the 1996 version. Similarly, were this so, the effects of modifying 1996 levels to include Mental Arithmetic would also have been enhanced.

How might standards in 1996 and 2000 differ?

Table 2.8.9 shows the 1996 and 2000 KS3 Mathematics threshold marks for each level, by tier. These are not directly comparable because of the inclusion of Mental Arithmetic in 2000. The hypothetical 'modified 1996 thresholds' (which mimic the incorporation of a Mental Arithmetic element equivalent to the 2000 version's - calculated as described earlier) are also included, as an aid to comparison.

Table 2.8.9 KS3 Mathematics cut-scores and equatings

	1996 cut-scores	*modified 1996 cut-scores, including MA*	*2000 cut-scores*	*2000 marks equated to 1996 cut-scores*
T3-5				
L2	20	31	27	*34*
L3	30	37	33	54
L4	61	77	69	88
L5	88	109	104	111
T4-6				
L3	24	31	27	?
L4	32	37	33	*53*
L5	54	66	60	69
L6	79	96	89	94
T5-7				
L4	22	35	32	?
L5	30	41	38	?
L6	48	65	61	74
L7	81	103	93	102
T6-8				
L5	22	41	37	*51*
L6	30	47	43	*56*
L7	54	76	68	67
L8	86	113	110	111

The 1996 cut-scores have been used as starting points for equipercentile equatings[5] to the 2000 test's total score scale and the resulting points on this scale, equated to the 1996 operational thresholds, are shown in the final column.

[5] Separate equatings for boys and girls were not feasible because of the relatively small numbers within tiers.

Note that data at the lower end of 2000 version mark ranges were sparse. Thus equatings for the lower thresholds in some tiers were impossible to estimate without extrapolating beyond the mark ranges observed (denoted by a question mark) or were potentially unstable because of gaps in the distributions (any involving the extreme 5% of a distribution are italicised).

Distributions of further hypothetical 'equated 2000 levels' which would have been awarded to pupils taking the 2000 version had the 'equated cut-scores' given above been used are shown in table 2.8.10. These are a good match to the distributions achieved by the group assigned to the 1996 version (shown in table 2.8.4), as would be expected given the equating method used.

Table 2.8.10 Distributions of hypothetical 'equated 2000 levels, by tier

	< L1	*L2*	*L3*	*L4*	*L5*	*L6*	*L7*	*L8*	
T3-5	1.0%	4.9%	41.2%	40.2%	12.7%				100%
T4-6			1.0%	10.1%	45.5%	43.4%			100%
T5-7				0.5%	4.2%	54.5%	40.8%		100%
T6-8	0.8%				1.5%	7.0%	60.5%	30.2%	100%

The experimental evidence suggests that the improvements since 1996, nationally, in the proportions of KS3 children reaching the highest levels (7 and 8) in their end of key stage Mathematics tests may be more defensible than the apparent improvements in Mathematics performance reported elsewhere in the ability range.

The equatings above suggest that in some instances substantial changes to the thresholds taken operationally in 2000 would have been required to bring them into line with the standards set for the 1996 version of KS3 Mathematics, especially for the lower levels.

- In T3-5, upward revisions of as many as 21 marks at the level 3 threshold, 19 marks at level 4, and 7 marks at level 5, would appear to be required.

- In T4-6, upward revisions of 20 marks at level 4, 9 marks at level 5 and 5 marks at level 6 might be called for.

- Even in T5-7, increases of 13 marks for the level 6 threshold and 9 marks at level 7 seem merited. Note that these data do not enable any estimates of the adjustments needed at lower levels in this tier.

- But in T6-8, whilst a case might be made for increasing the level 5 and 6 thresholds (by 14 and 13 respectively) the operational level 7 and 8 thresholds seem to require no adjustment.

Can we be confident in these equatings, given the relatively small numbers involved in some comparisons, once pupils were spread across the four tiers? Smaller numbers reduce the power for statistical tests to detect differencs, yet the size of effects here was such that statistical significance was not in doubt and the replicative pattern across tiers, with effects progressively reducing in size as ability rises, also adds to the credibility of these results.

2.9 KS3 Science experimental comparisons: 1996 v 2001

The historical pattern of KS3 Science test results since 1996 is shown in table 2.9.1 and indicates a substantial improvement. For instance whilst only 21% reached level 6 or better in 1996, 34% did so in 2001. A large improvement was evident in 1997 but it was not until 2000 that national results of a similar quality were reported again. 2001 saw further improvements.

Table 2.9.1 KS3 Science test results 1996 - 2001 (% of cohort at each level*)

	<3	*Level 3*	*Level 4*	*Level 5*	*Level 6*	*Level 7*	*Level 8*	*EP*
1996	2%	9%	26%	35%	17%	4%	0%	0%
1997	2%	8%	24%	31%	22%	7%	0%	0%
1998	4%	10%	25%	29%	20%	7%	0%	0%
1999	3%	9%	28%	31%	18%	5%	0%	0%
2000	3%	10%	23%	30%	23%	6%	1%	0%
2001	3%	7%	20%	32%	26%	7%	1%	0%

* rows do not total 100% as absentees and children disapplied are not shown

The validity of experimental comparisons in schools in Northern Ireland

Variations in style and content between the 1996 and 2001 versions

In both versions there are two tiers, targeted at levels 3 - 6 and 5 - 7. Within each tier there are two papers, paper 1 and paper 2, each one hour long, making the test a total of 2 hours for each tier. The questions in the tiers overlap, with some targeted at levels 5 and 6 being identical in both. Tier 3-6 carries 180 marks; tier 5-7 150.

Question style

Table 2.9.2 shows the number of marks for different item types in each tier/version of the 1996 and 2001 tests. There are fewer objective questions in both tiers in 2001 than in 1996. Instead, in 2001, these marks are divided between short answer and one word answer items, with the number of marks available for explanations and diagrams relatively similar in the two versions.

Table 2.9.2 marks available for different item types

*Item type**	*Objective*	*Diagram*	*One Word*	*Short Answer*	*Explain*
T3 - 6 1996	93	9	14	48	14
T3 - 6 2001	61	13	34	60	12
T5 - 7 1996	51	7	26	46	20
T5 - 7 2001	28	10	30	63	21

* Objective items involve multiple choice questions or tick boxes, or matching given information etc.
Diagram questions require drawing or adding labels etc. to a diagram
One word questions require one word answers
Short answer questions require brief phrase(s) or sentence(s)
Explanation questions require longer written answers and the command word in the item is 'explain'

Content

In the revised curriculum introduced in 2000 for England, aspects of science previously set out in the introduction to the Programmes of Study have been incorporated into Sc1 (scientific enquiry). Consequently, in contrast to the 1996 version, the 2001 test blueprint had been revised to include items

referring to experimental and investigative contexts targeted at a range of (rather than particular) levels. So in 1996 no marks were explicitly awarded for AT1 in either tier and, in both tiers, marks were more or less equally divided between ATs 2, 3 and 4. But in 2001 marks were awarded for ATs 1, 2, 3 and 4. The 2001 tier 3-6 paper had slightly more marks for AT2 than ATs 3 and 4, with AT1 having the smallest mark allocation (about half that of AT2). For 2001 tier 5-7 the marks are more evenly divided between ATs 2, 3 and 4, with a smaller proportion (about a quarter of that for other ATs) being awarded for AT1. But this overstates the shift in curricular emphasis, because even before this restructuring of the KS3 Science curriculum, KS3 Science national tests had deliberately incorporated questions which addressed aspects of Sc1. For instance, in the 1996 version of the test, T3-6 Paper questions 2, 4, 9 and 11 and Paper 2 questions 7, 8, 13, 14, and 15, and T5-7 Paper 1 questions 2 and 4 and Paper 2 questions 4, 5 and 6, all required pupils to interpret diagrams and tables of data or make statements about the procedures and safety aspects of experimental work etc.

The curriculum changes in England introduced in 2000 also included the removal of sexual reproduction of plants and the classification of solids, liquids and gases from key stage 3, so a small number of questions on these topics were included in the 1996 version but not in the 2001 test. However this amounts to no more than the usual variation in topics sampled each year.

Curricular and structural issues

Structural issues

KS3 spans the same age range (11-14) in England and NI, but whilst most children transfer from primary schools to secondary education at age 11 in both systems, NI has a largely 'selective' system post 11, in contrast with England where most children attend comprehensive schools. In the final year of KS2 in NI, transfer procedure tests are taken by those wishing to be considered for 'grammar' schools. Children who fail or do not take these attend 'secondary' schools. About 40% of NI children attend grammar schools, half of which are single sex schools. Most NI schools are also denominational, Protestant or Catholic, complicating provision still more, but such factors should not invalidate our experimental comparisons, provided that 'sampling' provides a reasonable balance of boys and girls across the ability range.

End of KS3 assessment arrangements in NI are similar to those in England, including externally set and marked formal tests in Science, although tiering arrangements for NI KS3 Science tests in 2001 were more complex than those in the QCA tests, with three tiers targeted at levels 3-4, 4-6 and 6-8. Nevertheless NI children should therefore be familiar with testing, including tests of the sort involved.

Curricular variations

The Programmes of Study for science in England and NI cover much the same areas and expected outcomes. Overall, the level descriptions cover the

same criteria, although some features appear at different levels. Aspects of the English science curriculum which do not feature in the NI curriculum documents are as follows:

AT2
- enzymes
- bacteria, viruses, immune system and medicines
- selective breeding - new varieties

AT3
- conservation of mass (physical change)
- conservation of mass (chemical change)
- geology

AT4
- transfer of energy from battery
- 'seeing'
- colour filters
- light travels faster than sound
- sun / stars as light sources
- satellites and probes
- electricity from variety of sources

Inspection of the 1996 and 2001 versions of the KS3 Science tests by project staff suggested that performance on a relatively small proportion of questions might be affected by these curricular variations between England and NI. The questions concerned and the resulting number of marks affected are shown in table 2.9.3 below.

Table 2.9.3 Questions affected by curricular disparities between England and NI

1996 Paper 1 *Tiers 3 - 6* ***marks***	*1996 Paper 2* *Tiers 3 - 6* ***marks***	*1996 Paper 1* *Tiers 5 - 7* ***marks***	*1996 Paper 2* *Tiers 5 - 7* ***marks***
-	Q13 geology **6** **6 / 180**	Q13 a) b) colours & light **4**	Q5 geology **6** **10 / 150**
2001 Paper 1 *Tiers 3 - 6* ***marks***	*2001 Paper 2* *Tiers 3 - 6* ***marks***	*2001 Paper 1* *Tiers 5 - 7* ***marks***	*2001 Paper 2* *Tiers 5 - 7* ***marks***
Q7a)c) electricity from variety of sources **4** Q9a)b)c) bacteria **4** Q13b) 'seeing' **3**	Q6 geology **6** Q15 c) I)ii) conservation of mass (chemical) **2** **19 / 180**	Q2a)b)c) bacteria **4** Q6b) 'seeing' **3** Q9b) colour filters **2**	Q6c)I)ii) conservation of mass (chemical) **2** Q12 geology **7** **18 / 150**

It would seem that relatively few marks on the 1996 version (only 6 / 180 on T3-6 and 10 / 150 on T5-7) were likely to be affected, but that performance on more marks (19 on T3-6 and 18 on T5-7) on the 2001 version might be inhibited if NI children have not in fact encountered the topics involved in their schools. Teachers in NI may in fact teach some of these topics and the Project's survey of reactions to the test materials by staff in the schools participating (see below) investigated this.

Teachers' opinions

Ten heads of science (from 5 schools administering T3-6 and 5 administering T5-7) of the fourteen schools administering the science tests completed the Project's questionnaire seeking their reactions. No difficulties with test administration were recorded, although one secondary school (administering

T3-6) commented on extensive pupil absences during the testing period. General comments were made by two secondary and three grammar schools. One secondary felt that some of the language in the lower tier test was not suitable for lower ability pupils. Two grammar schools commented that their pupils were suffering from 'exam exhaustion' and hence this may have affected their performance in the tests. But this would presumably have afflicted those taking both versions equally and hence not invalidated our experiment. One of these schools also commented that the type of questions used in the QCA tests were more straightforward than those in the NICCEA tests, so that although one school commented that the QCA tests were more difficult than the NICCEA tests, teachers do not seem to regard the QCA tests as problematic in any general sense.

Tier 3-6: All except one of the secondary schools administering tier 3-6 felt that both the 1996 and 2001 versions of the KS3 Science tests were either 'similar' to (n=2) or a 'very close' (n=2) reflection of the NI curriculum. The other school recorded a poor match for both versions, giving as their reason the view that the QCA tests were more difficult than the equivalent NI end of key stage tests[1]. When asked to describe specific difficulties their students might have with the tests, for either the 1996 and 2001 versions, two of the five schools commented that geology was not a part of the NI Programme of Study. When asked to pinpoint specific questions that could be affected by the above disparity, three of the five identified 1996 Paper 2 Q13 (6 marks) and the same number identified 2001 Paper 2 Q6 (6 marks) as potential problems. The teachers' view thus seems to be that both versions of this tier are little (and equally little) affected by curricular disparities between NI and England, although the project team's analysis of the curriculum documents and tests (above) had led to the view that the 2001 version was potentially rather more affected.

Tier 5-7: All five grammar schools administering tier 5-7 felt that both the 1996 and 2001 versions of the upper tier tests were 'similar' or a 'very close' reflection of the NI science curriculum. But when asked to describe any difficulties they felt their students might have with the tests, three schools commented that 'geology' was not a part of their Programme of Study. They considered various questions might be problematic and identified the following[2] as assessing topics they had not covered. Note however that the two (1996 P1 Q11 & P2 Q13) not picked up by the Project's desk analysis, reported above, do in fact test topics within NI curriculum documents.

1996 version

- Paper 1 Q11 (Compounds & the periodic table) 4 marks
- Paper 1 Q13 (How coloured objects appear in other colours of light) 4 marks
- Paper 2 Q5 (Geology) 6 marks
- Paper 2 Q13 (Blood = a transport medium; aerobic respiration & alcohol abuse) 8 marks

Total: 22 marks

[1] It is possible that this may reflect the narrower tiers in NI, which may make weaker pupils less likely to encounter questions beyond their capabilities.
[2] A record was taken of questions that were identified by 2 or more schools.

2001 version

- Paper 1 Q2 a, b, c only (Bacteria & viruses) 6 marks
- Paper 2 Q12 (Geology) 7 marks

Total: 13 marks

The teachers' view would seem to be that more marks in the 1996 version of T5-7 were affected by curricular disparities than was the 2001 version of this test. This conclusion is at variance with the Project's desk analysis, which identified additional 2001 questions (involving another 7 marks) as potentially affected.

Of course children will be able to attempt some (or parts of some) of these questions using knowledge gained elsewhere, and teaching in England might also not cover every topic. But to give due weight to teachers' opinions we should perhaps set aside the project staff's conclusion that performance on the 2001 version was likely to be the more inhibited by curricular variations. It could be safer to assume that the two versions are similarly affected.

The Data

The sample was designed to take account of the diversity of secondary schools in NI, with a good selection (rather than a random sample) of schools from throughout Northern Ireland being invited[3] to participate, including grammar and secondary schools etc. In all 14 schools agreed to assist the Project. Seven selective schools administered tier 5-7 of the tests (including one boys' school and two girls' schools), whilst seven secondary schools administered tier 3-6 (including 1 boys' school) to a total of 1552 pupils, using spiral allocation within gender to create two random groups. One group took the 1996 version of KS3 Science, whilst the other took the 2001 version.

One (mixed) school administering tier 5-7 only administered the 2001 version of the test and their data could therefore not contribute to our analyses. Checks on the numbers of pupils allocated to the two versions of the tests by each of the remaining schools, together with the means and variances of the scores recorded on each version within schools, suggested that they had managed to follow the quite complex instructions for allocating quasi-randomly matched groups successfully.

Completed test papers were returned to Cambridge for marking, together with the pupils' operational NI KS3 Science test results. A team of five experienced KS3 Science test markers, briefed and co-ordinated by a team leader with experience of operational marking in 1996 as well as 2001, each marked (randomly apportioned) scripts from each version and tier. After data entry and cleaning, statistical quality control analyses investigated variations in the marks awarded by different markers but no statistically significant differences between them were detected.

[3] Initial contacts with schools were made by NICCEA, on the behalf of the Project.

Some children were absent when one of the test papers was administered and schools were unable to provide details of performance by all pupils on their NI end of KS3 Science test, but full data were obtained for a total of 1366 pupils. Table 2.9.4 shows how these break down between versions, tiers and genders, revealing that those taking both versions of T5-7 included markedly more girls than boys - a result of the inclusion of two girls' schools and only one boys' school in the sample. The effect of gender imbalance is however the same with respect to the two versions of the test, which is the focus of this investigation. Subsequent analyses will also take gender effects into account.

Table 2.9.4 Numbers taking each version & tier, by gender

		female	*male*	*total*
Tier 3-6	*1996*	160 (50.3%)	158 (49.7%)	318 (100%)
	2001	168 (53.7%)	145 (46.3%)	313 (100%)
	total	328 (52.0%)	303 (48.0%)	631 (100%)
Tier 5-7	*1996*	225 (60.3%)	148 (39.7%)	373 (100%)
	2001	219 (60.5%)	143 (39.5%)	362 (100%)
	total	444 (60.4%)	291 (39.6%)	735 (100%)

n 1366

How well matched were the groups assigned to the 1996 and 2001 versions of the QCA KS3 Science tests? The distributions of NI KS3 Science test levels for the groups assigned to the 1996 and 2001 versions are shown in table 2.9.5. They are very similar, as would be expected from randomly assigned groups, and can be regarded as matched with respect to achievement on their operational NI KS3 Science tests.

Table 2.9.5 NI KS3 Science test levels (by version of QCA KS3 Sc taken)

	L2	*L3*	*L4*	*L5*	*L6*	*L7*	*L8*		*total*
1996		2.6%	10.1%	20.0%	37.6%	25.6%	4.1%	100%	n 691
2001	0.1%	2.5%	12.6%	21.8%	34.4%	24.7%	3.9%	100%	n 675
n 1366	*Chi-square Likelihood Ratio 4.93, df 6, n.s.*								

How do the results on the 1996 and 2001 versions of the test compare?

The distributions of overall test levels achieved via the two versions of the QCA KS3 Science test are shown in table 2.9.6. These too are quite similar, the chief difference being that a lower proportion of those assigned to the 2001 version were awarded level 5, with more assigned to both the adjacent levels - 4 and 6. Although a slightly lower proportion of those assigned to the 2001 version achieved level 7, it is not simply the case that results from one version are notably superior to those obtained via the other.

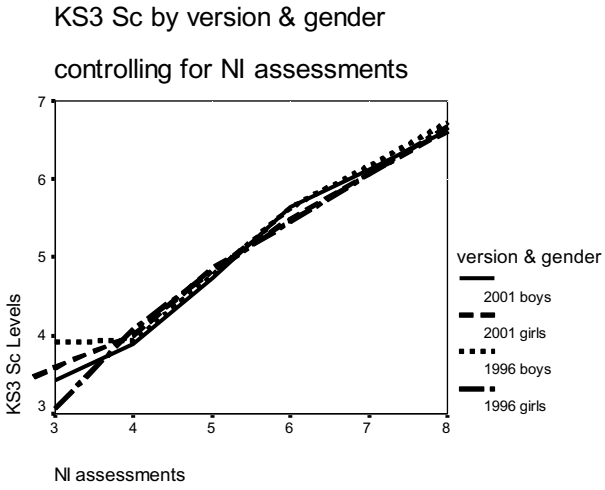
Table 2.9.6 QCA KS3 Science test levels (by version of QCA KS3 Sc taken)

	None	L2	L3	L4	L5	L6	L7		total
1996		0.9%	3.6%	11.1%	36.0%	38.5%	9.8%	100%	n 691
2001	0.4%	0.3%	4.6%	14.2%	32.0%	41.0%	7.4%	100%	n 675
n 1366	*Chi-square Likelihood Ratio 14.12, df 6, significance <0.05*								

Figure 2.9.1 investigates this issue graphically. It shows the linear local regressions (i.e. the average) QCA KS3 Science levels achieved by sub-groups of pupils (the groups of boys and girls assigned to the 1996 and 2001

versions of the test) awarded a given level in their operational NI end of KS3 test. Thus it shows average QCA KS3 Science test achievement, by version, whilst controlling for the ability in science of the sub-groups concerned.

Figure 2.9.1 Regression of KS3 Science on NI Sc test levels



The apparent disparities relating to the pupils achieving level 3 in their NI KS3 Science test should be ignored, as so few pupils scored at this level. Above this range, i.e. within the range of NI levels 4 to 8, figure 2.9.1 suggests that after controlling for variations in ability in Science between the sub-groups formed by gender and assignment to versions of the test, the levels awarded for achievement on the QCA KS3 Science tests were very similar; irrespective of gender or test version used.

Table 2.9.7 Mean & sd of QCA KS3 Sc test levels, by test version and gender

		mean level	*sd*	*n*
1996	*girls*	5.39	0.96	385
	boys	5.34	1.03	306
	all	5.37	0.99	691
2001	*girls*	5.33	1.04	387
	boys	5.25	1.04	288
	all	5.29	1.04	675
Total		5.33	1.02	1366

ANCOVA analysis
NI KS3 Science F = 1889.55, df 1, significance <0.001
Gender F = 4.10, df 1, significance <0.05
Version F = 0.18, df 1, n.s.
Interaction of Gender and Version F = 0.34, df 1, n.s.

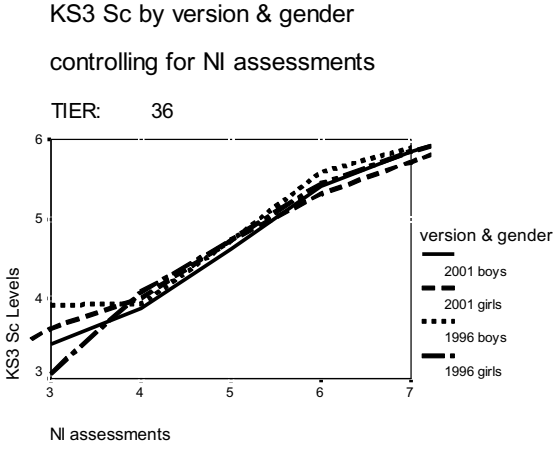
Table 2.9.7 presents the table of means and standard deviations of QCA test levels associated with figure 2.9.1, together with an analysis of covariance (ANCOVA), evaluating the statistical significance of differences in test levels obtained by those taking the 1996 and 2001 versions, together with the effects of gender, after controlling for NI KS3 Science levels. This confirms

that whilst gender differences (girls tending to obtain higher levels than boys in this sample) are significant, the effects of test version (i.e. 1996 v 2001) are not. Overall, it would seem that the standards implied by the cut scores selected for the two versions are much the same, so KS3 Science test standards appear stable over this time interval.

Despite reaching this overall conclusion, further similar analyses were carried out to explore test standards within the two tiers of KS3 Science. Figure 2.9.2 presents the linear local regressions of QCA levels achieved on NI test levels for each tier, whilst table 2.9.8 provides the associated tables of means by tier plus ANCOVA analyses evaluating the statistical significance of differences within each tier.

Figure 2.9.2 Regression of KS3 Science on NI Sc test levels, by tier

Tier 3-6



Tier 5-7

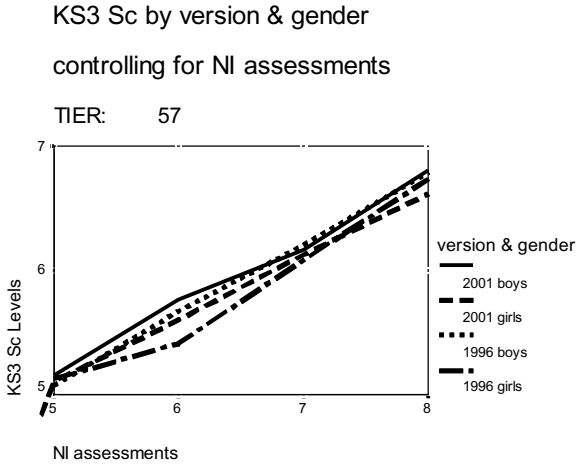


Table 2.9.8 Mean & sd of QCA KS3 Sc test levels, by test version, tier and gender

		mean level	*sd*	*n*
Tier 3-6				
1996	*girls*	4.83	0.96	160
	boys	4.82	0.99	158
	all	4.82	0.97	318
2001	*girls*	4.71	0.94	168
	boys	4.61	0.97	145
	all	4.67	0.95	313
Total		4.75	0.97	631
Tier 5-7				
1996	*girls*	5.80	0.73	225
	boys	5.91	0.72	148
	all	5.84	0.73	373
2001	*girls*	5.79	0.87	219
	boys	5.90	0.66	143
	all	5.83	0.78	362
Total		5.84	0.75	735

ANCOVA analyses

Tier 3-6	*NI KS3 Science F = 608.78, df 1, significance <0.001*
(n 631)	*Gender F = 0.01, df 1, n.s.*
	Version F = 3.75, df 1, significance 0.05
	Interaction of Gender and Version F = 0.27, df 1, n.s.
Tier 5-7	*NI KS3 Science F = 368.54, df 1, significance <0.001*
(n 735)	*Gender F = 10.60, df 1, significance 0.001*
	Version F = 0.72, df 1, n.s.
	Interaction of Gender and Version F = 0.02, df 1, n.s.

Within tier 3-6 it would seem that although the relationship between NI operational test levels and levels achieved via the QCA tests are similar for boys and girls, those allocated to the 1996 version in this experiment obtained slightly better outcomes than those allocated to the 2001 form of the test. After controlling for ability (via NI KS3 Science test results) the 1996 version of tier 3-6 appears to yield levels on average about a tenth of a level higher than the 2001 version. Though small, the difference is statistically significant.

Within the higher tier any variations in levels awarded from the two versions of the QCA KS3 Science test are not statistically significant. Whilst the graphical display of these data appears to suggest that of pupils achieving level 6 in NI Science tests, those taking the 2001 version of the QCA tests are on average awarded slightly higher levels, the difference between versions is not consistent across levels; closing or reversing at other points in the ability range.

Do standards in 1996 and 2001 differ?

The analyses within tiers in essence confirm the conclusion reached from analysing the data at an overall level.

Certainly the thresholds determined in 2001 for the higher (T5-7) tier measure up to the standards in the equivalent tier in the 1996 version of the test. In the lower tier (T3-6) these data might at face value suggest that it is even possible that the thresholds used in 2001 might have been a mark or so too high.

But given the modicum of disagreement between teachers and the project team concerning possible curricular effects, it is perhaps safest to conclude simply that this investigation gives us no reason to doubt that QCA have successfully maintained KS3 science test standards between 1996 and 2001.

3 Teachers' judgements of KS2 English scripts: 1996 v 1999 versions

Would teachers' professional judgements about the quality of work agree with our experimental findings? We mounted a small-scale exercise, using Northern Ireland children's KS2 English scripts from the 1996 and 1999 versions of the tests, to investigate this. A small group of teachers were asked to compare sets of scripts 'representing' key mark points identified by our empirical work and to discuss the issues arising.

Participants and programme

The ten teachers involved all had experience of teaching Year 6 and were recruited from schools in Cambridgeshire, Suffolk and Norfolk. Nine had not previously been involved in the development or marking of KS2 English tests. The tenth was an experienced KS2 English marker who was able to answer detailed questions about mark schemes if required. They agreed to attend a one-day event, in Autumn 1999, which included:

- Briefings from project staff about the Project, the structure of the KS2 English tests in 1996 and 1999 and the comparisons we wanted them to make.
- An opportunity to study the test materials and marking schemes and to ask questions about them.
- Individual reading and blind comparison of the sets of scripts representing key mark points.
- Individual recording of their blind comparisons between sets of scripts representing key mark points from the 1996 and 1999 versions, and the reasoning behind them.
- Group feedback from project staff about their judgemental comparisons.
- Group discussion about the issues arising.

The scripts and materials provided

Three sets of scripts (with five scripts in each set, taken from our experimental comparisons) were involved.

- Set 96-1/5, 'represented' the 1996 version level 4 threshold mark (i.e. 57)
- Set 99x-1/5, 'represented' the 1999 version level 4 threshold mark (i.e. 48)
- Set 99y-1/5, 'represented' the 1999 mark equated to the 1996 Level 4 threshold (i.e. 53)

Each set contained five scripts (i.e. the entire test completed by a single child, including Reading, Writing, Spelling and Handwriting elements) which were selected to represent the mark point concerned. Selection was a deliberate process, rather than at random, because of the need to exclude scripts which would render the judgements required impossible, rather than merely difficult. Selection of the scripts entailed:

- Identifying the pool of scripts available. First all children with total marks corresponding to each of the three key mark points were identified. There were fewer than five children at one of these, so the pool of scripts needed widening and children scoring one mark above and one mark below each of the key marks were also included.
- Listing each child's profile of marks on the Reading, Writing, Spelling and Handwriting elements that constituted the test.
- Screening out uneven profiles, where children reached their total by an unusually high mark on one element (say, Reading) to compensate for a poor performance on another (say, Writing), as these would have complicated comparisons.
- Selecting five scripts to represent each of the three key mark points with relatively even mark profiles and without much variation in the writing marks. In each case these included three scripts at the mark point itself, plus one at the mark above and one at the mark below.

The mark profiles of the selected scripts are shown in table 3.1.

Table 3.1 Details of selected scripts

	Script	*Reading*	*Writing*	*Spelling*	*Handwriting*	*Total*
1996 version Level 4 threshold	*96-1*	29	20	6	3	58
	96-2	27	21	7	2	57
	96-3	28	20	6	3	57
	96-4	28	20	6	3	57
	96-5	27	20	6	3	56
1999 version Level 4 threshold	*99x-1*	19	20	8	2	49
	99x-2	19	20	5	4	48
	99x-3	17	20	8	3	48
	99x-4	22	21	2	3	48
	99x-5	20	20	4	3	47
1999 version mark equated to 1996 Level 4 threshold	*99y-1*	24	20	7	3	54
	99y-2	21	21	8	3	53
	99y-3	21	21	8	3	53
	99y-4	24	20	5	4	53
	99y-5	24	20	6	2	52

Examiners' annotations, including total marks gained, were removed from each script and the scripts were labelled 96-1/5, 99x-1/5 and 99y-1/5, to identify the three sets whilst concealing the marks 'represented'. However, for the Reading and Spelling elements, the detailed marks showing whether the many individual questions within these elements were right or wrong were left in situ, to help teachers make sensible comparisons. We did not want to tempt the teachers to spend the time available reproducing the marking process or trying to apply the marking scheme without proper briefing or co-ordination.

In addition to photocopies of the scripts themselves the teachers were also provided with copies of the mark schemes for Writing, Handwriting and Spelling. The mark schemes for Reading were available for them to consult.

Comparative judgements

Following their initial briefing the teachers were asked to work individually, to scrutinise the scripts and compare them in terms of the overall 'quality of work' of the set. Forms were provided for the teachers to record their comparisons of the three sets of scripts.

They were instructed to begin by comparing the two sets of scripts from the 1999 version of KS2 English and to decide which set displayed the higher quality of work. They had the option of deciding that they could not see any difference between the two sets. They were warned not to spend too long on this phase and were then asked to consider the set of scripts from the 1996 version. They had to decide how the quality of these scripts (where children were responding to different stimuli and questions etc.) compared against the two sets from the 1999 version. In effect they were asked to rank the three sets of scripts.

We anticipated that the judges should have little difficulty with the Writing and Handwriting elements, as the criteria remained the same for children of this level of ability and we had selected scripts without too much variation in these respects. The most problematic feature in their task was to compare the Reading test performances, when the children taking the 1996 version had faced an inherently easier task. Spelling posed similar, though perhaps less extreme, problems.

The teachers were asked to tick one box in the columns for 99x-1/5 and 99y-1/5, in a grid much like that shown in figure 3.1, so as to compare them to the 96-1/5 set of scripts. The crosses shown in figure 3.1 indicate where we might hypothesise the judgements 'should' fall if the teachers agree with the equating from the experimental study. Teachers were also asked to write a brief rationale for their rankings and invited to make other comments.

Figure 3.1 'Anticipated' pattern of judgements

Results

The judgements made are shown in Table 3.2, where the ten teachers are represented as A to J and the judgements made by each are linked, to emphasise the pattern.

Only two of the ten teachers (A & B) thought that both the 1999 version Level 4 threshold scripts (99x-1/5:48) and the 1999 version equated threshold scripts (99y-1/5:53) were superior to the 1996 version Level 4 threshold scripts (96-1/5:57). One of these had reversed the two sets of 1999 version scripts; considering those at 48 marks better than those at 53. His was the only such reversal.

Another two of the ten teachers judged the 1999 Level 4 threshold scripts equal to the 1996 level 4 threshold scripts. One of these (C) thought the 1999 equated scripts better than the other two sets, whilst the other (D) could not distinguish between any of the three sets.

Table 3.2 Results of teachers' comparisons of sets of scripts

	BEST 99x-1/6	99y1/6
	A	B
	B	A C E
96- 1/6	C D	D F G H I
	E F G H I	J
	J	
	WORST	

However, six out of the ten considered that the 1999 Level 4 threshold scripts were not as good as those at the 1996 threshold.

One of these (judge E) took the view that the 1999 equated threshold scripts were rather better than the 1996 Level 4 threshold scripts. Four (F,G,H & I)

conformed to the anticipated pattern and judged the 1999 equated scripts equivalent to the 1996 level 4 threshold. The last (J) was less kind, feeling that both sets of 1999 scripts were worse than the batch representing the 1996 Level 4 threshold.

Comments and Discussion

The teachers were universally surprised when it was revealed that almost all the scripts involved had been awarded level 4. All had thought they were dealing with scripts from children within level 3.

When asked how they had set about their task and about the problems they had encountered, several agreed that they had tried to find questions testing similar things in the two versions of the test and concentrated on responses to these. Two scripts (99x-1 & 99x-2) were generally agreed to have been unusual, in that they appeared to come from relatively able children who had not completed the Reading test. But in general the selection of scripts seemed to have been adequate for the task.

Judge A explained his reversal of the two 1999 groups by suggesting that he found the responses of pupils with a mark of 48 'livelier' than those at a mark of 53, even though they were sometimes wrong or incomplete. Both judges suggesting that the two sets of 1999 scripts were superior made comments which suggested that their decision was heavily influenced by the view that the Reading element of the 1999 test was more challenging than the 1996 version. (e.g. Judge A - '… more comprehensive in testing language skills and more rigorous….lacking use of language styles and inferential questions..'; Judge B – '…1996 Reading was less challenging, as evident in unfinished and unanswered questions in 1999..'. Both (along with others) detected that the 1999 spellings were a little more difficult. Judge A commented that '… the 1999 Writing test was much fairer and gave much more scope and opportunities than 1996. This clearly affected the better writing results in the 1999 paper…'. Their comments perhaps suggested a focus on the questions, rather than the answers.

Many of the judges seemed to have discerned that there was little to choose between the Writing in the 1996 and 1999 level 4 threshold scripts. Those (the majority) who had ranked the 1999 threshold below the 1996 threshold based their judgement on the relatively poor answers from the 1999 version Reading scripts. Even though the 1999 version posed more challenging questions they concluded that, on balance, the quality of work did not match up to the 1996 scripts.

When project staff pointed out the implications of these rankings, namely that the majority view indicated that standards were more lenient in 1999 than in 1996, there was little resistance. One or two teachers found it hard to accept, pointing out that the 1999 test required a greater range of skills and was more demanding, especially for less able children. More generally, the teachers present liked the 1999 test, as had children in their classes who had taken it.

Before seeing the scripts they had perhaps not fully appreciated how challenging it would prove to be. But some then reflected on their results in 1999, recalling that their own teacher assessments had often proved less generous than their children's KS2 English test levels.

The layout and style of the1999 Reading test was often considered to be an improvement on the 1996 version (Judge A – 'accessible in the interest level and presentation'). But not everyone agreed. For instance Judge B thought the 1996 test 'was easier to follow and the questions more clearly set out at the top of each page.'

Conclusions

We should be careful not to rely too strongly on evidence like this, derived from the views of a small group of teachers who have looked at a handful of scripts. Given the imprecision of the judgemental process involved it was inevitable that a mixture of views would be recorded.

But the weight of opinion seemed to confirm the empirical evidence the project obtained elsewhere. The majority of the teachers involved considered that the work of children who (just) achieved Level 4 in 1999 was of lower quality than that of children (just) achieving Level 4 in 1996.

4 Children's perceptions of national test materials

4.1 Introduction & Methodology

The first pilot national tests in England were in 1992, so their history is relatively short. Since their introduction strenuous efforts have been made to improve the quality of the test materials and this drive for improvement has led to various changes in the character and format of the tests. Many of these have focussed on the desire to clarify the tasks involved, so as to help children demonstrate knowledge and skills they possess and to reduce the risk that tests fail to credit children when they have the abilities sought. In short, to make the tests as accessible as possible.

The Project's first phase included a strand of qualitative research developing new methodology to explore children's perceptions of the changes being introduced into national tests. The initial qualitative investigation looked at KS2 English Reading test materials and sought to provide evidence of children's appreciation of the features present in test materials (Green et al, 2001) and their reactions to differences between the 1996 and 1999 versions of the test. The underlying aim was to shed light on the possibility that the evolution of the test might have resulted in them becoming more user-friendly and conducive to the demonstration of achievement. The second and third phases of the Project's work broadened this qualitative approach by mounting similar investigations in the context of the other key stages/subjects: KS1 Reading, KS1 Mathematics, KS2 Mathematics and KS3 Mathematics in Phase 2 (where 1996 materials were compared with the 2000 versions) and KS2 Science, KS3 Science and KS3 English in Phase 3 (where the 1996 and 2001 versions were considered).

Methodology

Semi-structured interviews were designed to probe children's subjective experiences and perceptions and to investigate features of the tests which might facilitate or impede their performance.

Since the aim was to discover childrens' reactions to various styles and features present in test questions and associated stimulus materials from national tests for key stages 1, 2 and 3, we required a way to elicit personal responses from children as young as seven years of age. For interviewees aged 7, 11 or 14, verbalising thought processes is a challenging task and an interviewing method was designed to provide structure and support, so that interviewees were able to move gradually towards stating their personal preferences and giving reasons for them. To this end, Kelly's repertory grid questioning technique (Fransella & Bannister, 1977) was modified to take into account the age of the interviewees.

Kelly (1955) defines a construct as 'a way in which some things are alike and yet different from others'. Repertory grid technique is a process for eliciting personal constructs by presenting stimuli associated with the research

question and asking the interviewee to put two stimuli together and separate them from a third, giving reasons why the two are similar and the third is different. This was felt to be too complex for the children in this study and the approach was simplified. Children were shown only two stimuli and were asked to describe ways in which they were different, ways in which they were similar and, finally, which they would prefer to do and why. By concentrating on description initially, children were given the opportunity to consider the salient features of the test materials, simply describing what they could see. This enabled the children to consider the materials before attempting the more difficult task of stating preferences and giving reasons for them.

Selection of test materials for 'paired comparison'
For each key stage/subject the test materials for 1996 and either 1999, 2000 or 2001, as above, were carefully scrutinised to select materials which would ensure that the comparisons we asked children to make were as meaningful and productive as possible. In each case a series of pairs of questions/ sections were chosen which tested similar content and/or skills and levels, but which illustrated changes in the style of the tests etc. A wide range of attainment targets were represented and variations in question difficulty were minimised. The 'paired sections/questions' were presented to each child to target attention effectively. By sampling selectively from the tests in this way the amount of material under consideration at one time was manageable.

Samples
Small samples of children at the appropriate stage in their school careers were identified by their schools, with an even balance of gender and achievement in the subject concerned, as requested by the Project. The sample of 24 children initially used for KS2 English Reading was considered more than adequate, as interviewers encountered considerable repetition. Accordingly, to make effective use of the limited resources available for this strand of work, smaller samples (n 12) were used when work was extended to most[1] other subjects/key stages.

Interview schedules
The children familiarised themselves with the material, and were then interviewed using a semi-structured interview schedule. This allowed them to discuss salient features of the test materials.

Each interview schedule was developed to fit the needs of the particular comparisons being made. An example (for part of KS2 English) is reproduced as figure 4.1, to illustrate the instructions for interviewers and the approach taken and to provide an example of a form used to compare paired sections.

In each subject / key stage, pilot interviews with two pupils were carried out before the main study to refine the schedule and co-ordinate the approach of project staff involved.

Schedules each had an initial descriptive focus, supplemented with evaluative questions to probe children's preferences. Initial questions were sufficiently

[1] For KS3 Mathematics the complex tiering stucture required a modest increase, to n = 16.

open-ended to allow for ‘unexpected responses’, enabling the children themselves to identify salient elements.

The aim of the initial questions was to encourage description of the separate sections in the interviewees’ own language. Children’s responses identified the salient features, which were probed as far as possible. So, for example, if a child suggested that the sections were different because ‘*Section A is easier to read’*, this would be explored by asking, ‘*What do you think makes Section A easier to read?’* The response might then be, *‘The writing is different.’* This might be followed by, ‘*Can you describe the writing in Section A? ..... Now describe the writing in Section B.’* By allowing the child to identify the feature it was often possible to take the descriptive process further. If children had problems identifying features, they were directed towards aspects of the material, without being too restricted or too firmly led by the interviewer.

After probing for salient features, children were asked to express their preferences, via prompts such as: ‘*Which of these two sections would you rather do? Why?’* General answers such as, *‘I would rather do this one because it’s more interesting,’* were probed further. *‘What do you think makes this more interesting?’* Probes here related to ‘liking’ rather than ‘difficulty’.

This process was repeated for each of the paired sections with each child, although the sequence of administration of the ‘paired sections’ was varied to balance order effects. Interviews were taped.

Data analysis
The three researchers who conducted the interviews also analysed children’s responses to determine features of the test materials which had been identified. The analyses from their sets of interviews were then merged and responses were categorised. The reliability of categorisation was ensured by involving all members of the interviewing team. Frequency counts were undertaken and the effects of gender and level of ability were considered. Certain features were common to all paired sections under discussion and these are reported, for each subject/key stage, as ‘General Features’. Other features were specific to certain sections or to individual questions, or groups of questions. These are reported as ‘Section/Question Specific Features’. In phases two and three of the project, two 'warm up' questions were introduced to elicit children's opinions about 'test taking' and to investigate their recollections about personal feelings before and after their operational tests had taken place. Their responses are reported under the heading 'views on testing'.

The reports for each subject/key stage which follow, in each case, include details of the selected 'paired questions/sections' which provided the basis for interviewing as well as accounts of the salient features identified, preferences expressed, and the issues raised.

They are presented by key stage and subject, ordered as below, and each can be read independently, so that readers may select particular subjects or key stages if they prefer.

KS1 Reading
KS1 Mathematics

KS2 English: Reading
KS2 Mathematics
KS2 Science

KS3 English
KS3 Mathematics
KS3 Science

Children's views provide insights on the tests and the ways in which we have sought to 'improve' them. Their views are seldom heard in these professional debates and they have things to say which carry interesting implications for test developers (who will no doubt continue to seek improvements) and for all those concerned with the validity of national tests and the standards of achievement they represent.

Figure 4.1.1 Illustrative extracts from KS2 English Interview Schedule

KS2 English: Instructions for Interviewers

The following three phase interviewing process should be repeated for each paired section, focussing first on salient features identified in the stimuli etc., second on features of the questions themselves and third on preference (and the reasons for it) for one version or the other. Record the child's views on the forms provided for each.

1 STIMULUS MATERIALS

a How are these sections different?
Probe to encourage independent description of both sections based on salient features identified by the child.

b How are they alike?
Probe to encourage independent description of both sections based on salient features identified by the child

e.g. Child: 'This one is set out better.'
Interviewer: 'Now look at section A, can you tell me how it is set out?'
'What about section B, how is it set out?'

c *Introduce prompts to cover features to be targeted, as shown on the recording forms for each pair of sections.*

2 QUESTIONS

a Repeat 1 above, this time focussing on the questions related to the stimuli.

3 PREFERENCES

a Which section would you rather do?
b Why?

The aim is to encourage evaluative comparisons and to probe reasons for preferences.

Children should be encouraged to expand on general answers e.g.

Child: 'I would rather do this one because it's more interesting.'
Interviewer: 'What do you think makes section A more interesting?'

Introduce prompts to cover features to be targeted, as shown on the recording forms for each pair of sections

Please tape interviews and take notes on the recording forms

KS2 En Recording Form: A1 (section 1/ 96) A2 (section 1/ 99)

Name ..... B / G Reading TA Level .....

1 STIMULUS MATERIALS

1 How are these sections different? Probe responses re Stimuli

2 How are these sections alike? Probe responses re Stimuli

3 Prompts Stimuli

Content:	What is it telling you about?
Pictures / Diagrams:	What about the pictures / diagrams?
Layout:	What about the way it's set out, ... headings, ... print?

2 QUESTIONS

4 How are these sections different? Probe responses re Questions.

5 How are these sections alike? Probe responses re Questions.

6 Prompts Questions

What about the kinds of questions?
What about the ways of answering?
Q9 page given / arrows
Marks shown

3 PREFERENCES

7 Which section would you rather do?

8 Why? Probe responses
Look at the section, can you tell me why it's ..... more interesting/ easier to follow/ funnier?
Now, what makes the other section less interesting/ difficult / funny?
What about the questions?

4.2 KS1 English

Materials from the KS1 Reading Comprehension tests for 1996 and 2000 were selected so as to form pairs of questions/stimuli to serve as the basis for the interviews. These selections are described in table 4.2.1, which includes brief details of the features providing the focus for each 'paired' comparison.

Table 4.2.1 KS1 English qualitative comparisons – paired questions/sections

	1996	*2000*	*Focus of Comparison*
A	*A1* Front cover Page 1 'The Surprise' title page/child details	*A2* Front cover/child details 'Useful words page' Contents Page	● Initial appeal ● Pictures ● Layout child details ● Use of contents and useful words pages
B	*B1* 'The Surprise' narrative (without questions) Whole story read and pages 3;4 and 9 viewed	*B2* 'Mr Davies and the baby' narrative (without questions) Whole story read & pages 2,4 and 8 viewed	● Content ● Pictures ● Layout of story
C	*C1* Questions 11& 12 (on story)	*C2* Questions 7& 8 (on story)	● Use of bold ● Question types ● Spacing ● Font size
D	*D1* 'Riding on Trains' (informative text) page 15	*D2* 'Dogs' (poem)	● Content ● Text type ● Pictures
E	*E1* 'Danger' (informative text) page 16	*E2* 'What to do when you meet a dog' (informative text) page 14	● Content ● Text layout (bullets/spacing)
F	*F1* Question 12 (on info text)	*F2* Question 25 (on story)	● Question types ● Stem v response length
G	*G1* Question 13 (on info text)	*G2* Question 5 (on story)	● Question types

A Cambridgeshire school allowed the Project access to Year 3 children during the first half of the autumn term 2000. Due to the Project's schedule, we interviewed Year 3 children at the beginning of the school year about their experience of the national tests in the previous term. A sample of 12 children were 'selected', with an even balance of gender and teacher assessed levels in reading attainment, as shown in Table 4.2.2 below. The children were given an opportunity to familiarise themselves with the selected material before being interviewed.

Table 4.2.2 Interviewing 'Sample'

	Reading: Teacher Assessed Levels		*Total*
	Level 2	*Level 3*	
N boys	3	3	6
N girls	3	3	6
N total	6	6	12

In pointing out similarities and differences, children referred to a range of features. Note that they were not asked to state preferences at this stage, but merely to describe the materials.

General Features

Stimulus Materials
The following features of the stimulus materials were mentioned, in descending order of frequency:

- content — focus of the text:- characters, plot or theme.
- pictures — appearance, size, colour and quantity.
- title — size
- text length — more/less reading required in a particular section

Questions
The following features of the questions were mentioned in descending order of frequency:

- variety of questions — ways of answering
- font — selective use of bold, font size
- demand — amount of writing or reading required
- layout — spacing, position of questions, style of response lines.
- mark-boxes — presence in 2000 materials, shape

Section Specific Features

A1 1996 front cover & child details / A2 2000 front cover & reading aids

- reading aids - presence in 2000 materials of 'useful words' and 'contents' page.
- child details - which page they featured on:- Page 1 (1996) & front cover (2000)
- speech bubbles - on 'useful words page' in 2000 test.

B1 1996 narrative / B2 2000 narrative

- presence of author's name on opening page of the 2000 narrative
- different positioning of text and pictures (in relation to one another) in the two versions
- opening sentence of 2000 narrative: 'Once upon a time….'

D1 1996 'Trains' (informative text) / D2 2000 'Dogs (poem)'

- rhyming words in the poem 'Dogs'
- instruction in 2000 text: ie poem is read by the teacher.

E1 1996 'Danger' (informative text) / E2 2000 'What to do when you meet a dog' (informative text)

- bullet points - shape, ie circles in 1996 and dogs in 2000

Preferences

After considering the materials, children were asked which of the paired questions/sections they would prefer to do and, having expressed a preference, were asked to explain why. They were asked to comment on the stimuli and/or questions concerned and their replies often referred to the

features they had discussed earlier. Children were able to comment on the features which they felt helped them engage with the questions/texts etc.

A1 1996 front cover & child details / A2 2000 front cover & reading aids
Asked to compare the cover pages, almost all (11 of the 12) preferred the 2000 version (A2).

Reasons for preferring 2000
- initial appeal - dogs/animals were 'more interesting' or liked
- the pictures were considered 'funny'

B1 1996 narrative text / B2 2000 narrative text
Preferences were divided equally between these narratives (6 to each version). Most boys (n=4) preferred B1, the 1996 story, '*The Surprise'*, which involved a surprise train-ride for a group of friends on their school holidays. In contrast, most girls (n=4) preferred B2 *'Mr Davies and the Baby'* (2000) which describes how Mr Davies (a very excitable dog) eventually manages to go out walking with a baby and its mother. Ability also seemed to be a factor. Most level 2 children (2 boys & 2 girls) preferred *'The Surprise'*, whilst most level 3 children (3 girls and 1 boy) preferred the 2000 narrative.

Reasons for preferring 1996
- content - more characters made it more fun/ like surprises/ characters ask questions

Reasons for preferring 2000
- content - more detail about characters/ funny
- pictures - 'give you a bigger clue about what might happen'/ funny

C1 Q11 & Q12 (on story) v C2 Q7 & 8 (on story)
A slight majority (7 to 5) preferred C1 (1996 questions 11 & 12, p8). These questions included a multiple-choice question and one asking for a single sentence response. The 2000 questions (C2 - questions 7& 8) were similar in type, but required more reading. Interestingly most boys (n=4), including all those assessed at level 2 (n=3), preferred the 2000 questions. The format of the 2000 mcq was also different, in that the principle word was in bold font, as opposed to the whole of each possible response, as had been the case in the 1996 version.

Reasons for preferring 1996
- font - bold letters clearer/more interesting

Reasons for preferring 2000
- answers needed aren't very long
- looks harder and you learn more
- looks easier

D1 1996 'Trains' (informative text) / D2 2000 'Dogs (poem)'
The majority (by 9 to 3) preferred the 2000 questions (D2, relating to 'Dogs' - a poem) to the 1996 selection (D1, 'Riding on trains', an informative text).

Reasons for preferring 1996
- content - trains are fun/ learn something new
- pictures - more detailed

Reasons for preferring 2000
- content - like dogs
- pictures - 'nice' and colourful
- text type - poem more fun to read/ like rhyming words

E1 1996 'Danger' (informative text) / E2 2000 'What to do when you meet a dog' (informative text)
More (n 8) preferred the 2000 materials (E2, 'What to do when you meet a dog' - informative text) to the 1996 version (E1, 'Danger' - informative text) (n 4). Three of the four choosing the 1996 text were level 2 children, whilst most at level 3 preferred the 2000 version.

Reasons for preferring 1996
- content - like trains/ 'like to know about olden days'
- pictures

Reasons for preferring 2000
- content - you learn something
- pictures - brighter and more colourful
- text length - less reading

F1 'Danger' v F2 'what to do when you meet a dog' (informative texts)
Almost all (11 of 12) preferred F2 (2000 Q25) to F1 (1996 Q12).

Reasons for preferring 1996
- 'I like writing'

Reasons for preferring 2000
- question type - looks easier and quicker/ just choose and tick/ don't have to write/ lots of ideas, 'if you have trouble you can just pick one'

G1 Q13 (on informative text) v G2 Q5 (on story)
G2 (2000 Q5) was preferred by 9 out of the 12 children over G1 (1996 Q13).

Reasons for preferring 1996
- content - trains
- question type - easier

Reasons for preferring 2000
- question type - easier/ first answer given
- demand - 'just need to write two words'

Thus, overall, the majority of the children clearly preferred the selected 2000 test materials to the 1996 versions with which they were paired. There were two exceptions; in one instance (C1 v C2), more children preferred the 1996 version, whilst for B1 v B2, the narrative sections, preferences were evenly split.

Considering choices by individual children, only one (level 2) boy appeared to prefer the 1996 sections sampled overall. The remaining five boys and six girls chose more of the 2000 sections than the 1996 versions, although in four cases the margin was only one section.

Gender did not seem to be an issue, as the patterns observed for boys and girls were similar. In terms of ability, the preference for the 2000 sections was slightly stronger amongst the abler (level 3) children.

Views on test taking

When asked *'What do you think about children being asked to do the tests?'*, all of the children responded positively; commenting that they felt it was a good idea. Seven of the children went on to comment that they thought the purpose of the tests was to enable them to learn more.

But despite the above comments, when asked *'How did you feel about doing the tests?'*, only three children (all level 3) commented positively on their feelings *before* being tested. Their answers ranged from feeling *'alright'* before the test, to feeling *'excited, because my sister said it would be fun'.* Four children made no comment, and the remaining five children all said they felt either scared or nervous, mainly thinking that they would not be able to do the tests because they might be too hard. *After* testing two of the children felt relieved and glad it was over; a further five commented that they felt 'happy' or 'proud', especially, as the experience had not been as bad as they thought it initially would be. Five children made no comment.

Summary

- Overall, children favoured the questions/stimuli selected for these comparisons from the 2000 version of the KS1 Reading Comprehension test to those selected from the 1996 version. The 2000 versions of five of the seven paired selections compared were preferred.

- As far as the stimulus materials were concerned, certain features were often mentioned during the description of similarities and differences between the two tests. These included: the content of the text, in terms of the characters, plot or theme; the appearance, size and number of pictures in the text; the length of the text. As far as the questions were concerned, the main features were the variety of question types (which involved different ways of answering); the selective use of bold; varying font sizes; and the amount of reading and writing involved in the questions.

- The features described above were also frequently referred to when the children stated the reasons for their preferences. However, descriptive features which were section-specific (eg:shape of bullet points in section E) were rarely mentioned at this stage.

- In terms of the stimulus materials, the expression of preferences highlighted how highly the children rated content. They largely preferred texts that had characters they could relate to, had an interest in, or generally just 'liked'. Perhaps, unsurprisingly, the word 'fun' was used widely to describe what children found appealing about the characters or theme of the text. It was very clear that the accompanying pictures were also an important feature which helped the children engage with the text. They favoured pictures which were funny, detailed, bright and colourful.

- Children's preferences revealed that ‘poems’ were particularly popular. The children felt that this genre was fun to read and that they enjoyed the rhyming structure of the poem. In respect of the information texts sampled, the children frequently commented that alongside the appeal of pictures, they liked texts that enabled them to learn, or find out something new. This reasoning seems to fit well with the majority of the children’s ideas about the purpose of test-taking - ie *“so you can learn”.*

- In terms of the questions, their preferences revealed that the children favoured questions which were easy, required a quick response and involved little, or no writing. Multiple-choice items were therefore popular, (eg:F2) or questions which required only a one or two word response (eg:G2). The use of large and bold font was also mentioned with the children commenting that such features made the question clearer, more interesting and easier to see.

- Gender appeared to be an issue with reference to the narrative texts. The majority of the girls favoured the story of *‘Mr Davies and the Baby’* (2000), making reference to the text being ‘fun’ and more detailed and the pictures providing better predictive cues. Again the pictures were also described as ‘funny’. The majority of the boys, however, preferred the alternative offered, *‘The Surprise’* (1996), making reference to liking the pictures of the train and the number of different characters in the story.

- When levels of reading attainment were considered, the children’s preferences again differed in respect of the narrative section. The majority of level 2 children preferred the 1996 text *‘The Surprise’*. Their reasons mainly focussed on the content of the story (ie liking surprises and the characters). However, the majority of level 3 children preferred the alternative text from the 2000 test, *‘Mr Davies and the Baby’*. Their reasons were more far-ranging including the appeal of the content, namely, ‘Mr Davies’, alongside enjoyment of the humorous pictures and the greater level of detail in the story.

- All of the children responded positively when questioned on their views about being asked to do the tests. All thought that they were a good idea, with over half of the children also commenting that they thought that the purpose of the tests was to encourage them to learn more.

- However when the children described how they felt before they took the tests only three commented positively, a further four made no comment and the remaining five were nervous or scared due to fears about the difficulty of the tests. When describing how they felt after the test, the children were again more positive. Five were happy and proud as the experience had not been as bad as they had anticipated, whereas two were just relieved that it was over. The remaining five children made no comment. It would seem that ‘fear of the unknown’ or ‘fear of the difficulty of the unknown’ was the main cause of anxiety.

4.3 KS1 Mathematics

The pairs of questions from the 1996 and 2000 KS1 mathematics tests selected to illustrate the ways in which the test has evolved are shown in table 4.3.1. The selections sought to include a range of attainment targets and levels of difficulty as well as various types of question. The aim was also to avoid the repetition of question types with the same features as this would have led to repetition and could have been frustrating for interviewees.

Table 4.3.1 KS1 Mathematics Test - paired Questions

NC domain	*Level 2*	*Level 3*	*Focus of comparison*
understanding and using measures		Q23 1996 Q5 2000	• oral vs. written question
position and movement	Q29 1996 Q31 2000		• how question set out • grid vs. no grid
patterns of property and shape	Q9 1996 Q7 2000		• style of diagrams
classifying, representing and interpreting data		Q11 & 12 1996 Q10 & 11 2000	• way information shown • ways of answering
solving numerical problems	Q18 1996 Q32 2000		• context vs. no context
relationships between numbers and developing methods of computation	Q3 1996 Q9 2000		• ways of answering
developing an understanding of place value	Q22 1996 Q29 2000		• way set out

A Cambridgeshire school allowed the Project access to Year 3 children during the first half of the autumn term 2000. Due to the Project's schedule, we interviewed Year 3 children at the beginning of the school year about their experience of the national tests in the previous term. A sample of 12 children were 'selected', with an even balance of gender and teacher assessed levels in mathematics attainment, as shown in table 4.3.2. The children familiarised themselves with the questions, and were then interviewed according to a semi-structured interview schedule.

Table 4.3.2 Interviewing 'Sample'

	Mathematics: Teacher Assessed Levels		*total*
	Level 2	Level 3	
n boys	3	3	6
n girls	3	3	6
n total	6	6	12

General Features

The following features were mentioned, in descending order of frequency:

- context the 'story' within which the mathematics is set

- response types e.g. matching line, ticks and crosses, circling answers
- demand what they are asked to do, e.g. find shapes, which holds most
- operation whether question involves +, -, x
- writing amount of writing in the question
- box for answer where to write answer
- bold font compared to lighter font in writing and pictures
- mark boxes presence in margin, different shapes
- spacing amount of white space on the page
- position of writing how writing is positioned relative to other features, e.g. diagrams
- font size of writing

Question Specific Features

Clocks 1996 Q23 / 2000 Q5

- the hands on the clocks
- number of clocks on page
- no instructions on Q5 compared to Q23 which shows you what to do
- teacher reads it

Position and Movement 1996 Q29 / 2000 Q31

- grid / squares on Q31 compared to just plain and 'floating in the air' on Q29
- about symmetry / right angles
- use a mirror
- 'normal' shapes on Q29 compared to unfamiliar shapes on Q31

Properties of shapes 1996 Q9 / 2000 Q7

- plain black shapes in Q7 compared to shapes with pictures and writing on, looking like boxes of chocolates etc. in Q9
- names of shapes listed down the side in Q7

Data Handling 1996 Q11 & 12 / 2000 Q10 & 11

- numbers on graph's axis on Qs 10 & 11
- table in Qs 11 & 12 compared to graph in Qs 10 & 11

Numerical Problem 1996 Q18 / 2000 Q32

- picture of girl and food, helps you to remember, gives you ideas what things are
- nothing explained in Q32, just a take away sign, but Q18 explains it more, more like real life

Developing methods of computation 1996 Q3 / 2000 Q9

- circles with numbers in Q9
- diamonds and triangles in Q3
- answer given in Q3
- different totals in sums - 12 in Q3 and 17 in Q9

Place value 1996 Q22 / 2000 Q29

- about quarters and halves
- a number pattern in Q22
- arrows in Q22

Preferences

After they had considered the questions, children were asked which of each pair they would prefer to do and to explain why.

1996 Q23 v 2000 Q5

Preferences were equally divided between the 1996 (Q23) and 2000 (Q5) questions (6 v 6).

Reasons for preferring 1996

- looks easier
- won't forget instructions
- hands already drawn

Reasons for preferring 2000

- easier
- like drawing hands
- less on page
- looks fun/different

1996 Q29 v 2000 Q31

Two-thirds of the children (8) preferred the layout with a grid provided in B2 (2000 Q31) to the 1996 Q29 which did not have one, including five of the six girls and five of the six level 3 children.

Reasons for preferring 1996

- easy
- good at symmetry
- only one thing to do

Reasons for preferring 2000

- interesting
- easier/ better explanation and easier words
- less to write
- mirrors (on Q29) are fiddly

1996 Q9 v 2000 Q7

A slight majority (7 to 5) preferred C2's (2000 Q7) plain shapes to C1's (1996 Q9) decorations and words, including five of the six boys and five of the six level 3 children.

Reasons for preferring 1996

- looks easier
- easier to read
- funny pictures - nice to look at
- less to do

Reasons for preferring 2000

- looks easy
- know the answer
- like matching questions

1996 Q11 & 12 v 2000 Q10 & 11

The majority (8 of 12) preferred D2 (2000 Q10 & 11), in table form, to the graph in D1 (1996 Q11 & 12), especially level 2 children (5 of the 6).

Reasons for preferring 1996

- easy/ less complicated
- looks better

Reasons for preferring 2000
- interesting because of pictures
- less / little to write
- looks easier
- looks fun

1996 Q18 v 2000 Q2000

More (n 7, including 5 of the 6 girls) preferred E1 (1996 Q18), which was contextualised and included supporting illustrations, to E2 (2000 Q32) (n 5).

Reasons for preferring 1996
- interesting
- nice pictures
- easy words/ easier or lower numbers
- might need it in real life
- takes less time
- less to work out in your head

Reasons for preferring 2000
- good at take away
- looks easier
- not much to do
- takes less time
- prefer sum to lots of words

1996 Q3 v 2000 Q9

More (n 7, including 5 of the 6 girls) preferred F2 (2000 Q9), which required children to tick the two numbers making a total, as opposed to writing them in, as in F1 (1996 Q3) (n 5)

Reasons for preferring 1996
- smaller numbers
- easy/ easier numbers
- explains it better
- layout

Reasons for preferring 2000
- looks easier
- only small numbers
- less boring

1996 Q22 v 2000 Q29

A majority (n 8, including 5 of the six girls and five of the six level 3 children) preferred G2 (2000 Q29), which set the fractions question in a 'chocolate' context, to G1 (1996 Q22) (n 4).

Reasons for preferring 1996
- chocolate theme
- easy/ looks easier
- only have to do it once
- clearer instructions
- like quarters

Reasons for preferring 2000
- easy

- good at number patterns/halves
- just put in answer

Of the 7 pairs of questions sampled more of the children preferred the 2000 questions overall. Looking at the totals of all the children's choices, in only 1 case did more prefer the 1996 question, whilst for another pair of questions the preferences were evenly split.

Considering the individual children's choices, 2 boys and 2 girls preferred the 1996 questions overall, while 4 boys and 4 girls chose more of the 2000 questions. However, for 7 of the children the difference was only one pair.

Boys *and* girls preferred the 2000 questions overall. However, the preference for the 2000 questions was strongest among the level 3 children. But the preferences expressed by level 2 children showed similar overall patterns to those of the whole sample.

Views on test taking

When the children were asked *'What do you think about children being asked to do the tests?'* 3 of the 12 responded positively, commenting that it was exciting and good fun. Half (6) responded negatively, commenting that they felt scared, that they were boring or really hard or weird or strange. Three of the children made no comment.

When *asked about their feelings* before their operational KS1 tests, one level 3 boy commented that he felt excited and another said that he felt okay. The other 10 children commented negatively, saying they felt nervous or scared. More than half of their concerns stemmed from worries about getting questions wrong.

After the tests, 5 of the children felt relieved that they were over. They referred to the fact that they wouldn't have to repeat the experience, which resulted in 5 of the children feeling better, happy, relieved or good. They expressed mixed opinions as to whether they'd found the tests easy or difficult.

Summary

- The questions compared were only a small selection from the 2 versions of the test, but overall the 2000 version questions were considered the most appealing.

- The main features of questions noticed by children were their context; response types; what the question actually asked (e.g.How many...?); the type of operation involved in achieving a solution; and the amount of writing in the question.

- Features which were frequently mentioned when similarities and differences were described were rarely reflected in the children's reasons for their preferences.

- The reasons given for children's preferences most often related to the question being easy or looking easy. This reason was given between 3 and 9 times for every pair of questions. When probed, this often meant familiarity either with the layout or, more often, the content of the question. Other factors noted were if children knew the answer or were good at the particular content area. Children also commented on the words being easier or the numbers themselves, with smaller numbers equating to easier numbers.

- Often (more than half the paired selections) references were made to the amount that needed doing in order to answer the question. The children favoured those with less to do or write and those which took less time.

- The other characteristic which was commented on frequently was the appearance of the question. Children commented on questions looking fun or different and on pictures looking nice, funny and interesting.

- Girls showed more agreement in their preferences than the boys. Girls favoured questions with grids, ticking the answers and those where context and supporting illustrations were present.

- The level 3 children also favoured questions with grids and context, but differed from the girls by favouring plain shapes as compared to decorated ones. The level 3 children's preferences for the 2000 questions were more marked than the level 2 children, where the choices were more evenly split.

- There were mixed responses when the children were questioned about being asked to do the tests. The researchers felt that this was affected by the fact that these very young children did not always fully comprehend the question. Before completing the tests, most of the children had experienced negative feelings, being nervous and scared and worrying about answering incorrectly. This seems to link well with their reasons for their preferences which refer primarily to the ease of a question. They also reported on their feelings of relief/ happiness after the tests and to getting the tests over with.

4.4 KS2 English Reading

The two versions of the Reading tests were examined and materials were selected from them to form 'paired sections', each containing 'equivalent' 1996 and 1999 test materials, as detailed in table 4.4.1 below. These paired sections (A1 & A2, B1 & B2, C1 & C2) were presented to each child, so that in due course they considered both versions of the test. Sections were paired so that where possible similar types of text were compared.

Table 4.4.1 KS2 English Reading Test - Paired Sections

1996		**1999**	
A1 Cover Contents Instructional Text / Diagram	Section 1 Questions 4 & 5	**A2** Cover/Contents Introduction Explanatory Text / Captions	Section 1 Questions 1 – 9
B1 Information / Reference Text	Section 1 Questions 1, 2 & 3	**B2** Information / Reference Text / Cartoons	Section 3 Questions 1 – 7
C1 Narrative Text	Section 2 Questions 1 - 20	**C2** Poem	Section 2 Questions 1 – 9 Section 4 Questions 1 & 2

Two Cambridgeshire schools allowed the Project access to Year 6 children, in the first term of the 1999/2000 school year. A sample of 24 children were 'selected', half from each school, with an even balance of gender and teacher assessed levels in reading attainment, as shown in table 4.4.2.The children familiarised themselves with the material, and were then interviewed using an open-ended schedule. This allowed them to discuss salient features.

Table 4.4.2 Interviewing 'Sample'

	Reading: Teacher Assessed Levels			
	Level 3	*Level 4*	*Level 5*	
n boys	4	4	4	12
n girls	4	4	4	12
n total	8	8	8	24

Part of the KS2 English interview schedule is reproduced in 4.1 above, where it serves as an example of the general form of interviews and provides an example (for A1 & A2) of the form used to compare the paired sections.

Children pointed out a range of similarities and differences in the materials.

General Features

The following features were mentioned in descending order of frequency:

Stimulus materials

- general appearance - borders, bullets, stars

- length of text - problems of reading time, difficulties in retrieving information (mainly L3)
- font size
- bold print
- eye-catching colours

Questions
- multiple choice questions
- 'straightforward' questions - this term was frequently used to describe a type of question where one or two lines were given for the answer to be written.
- variety of questions - different ways of answering
- long questions
- more space for answers
- 'opinion' questions

Section Specific Features

A1 (1996 Cover etc & S1/Q4, 5) and A2 (1999 Cover etc. & S1/Q1-9)
- interest level of content
- supporting introduction and contents
- stimulus page shown in question - Q.9 p.6 1999
- marks for questions - preferred in right margin (prompted only)

B1 (1996 S1/Q1,2,3) and B2 (1999 S3/Q1-7)
- cartoons, humour
- serious, factual content in stimulus
- pictures included in question page
- headings asking questions
- familiarity of content

B1 (1996 S1/Q1,2,3) and B2 (1999 S3/Q1-7)
- descriptive content of poem

Preferences
Children were asked which of the paired sections they would prefer to do and why. They were asked to consider the stimuli and the questions before stating a preference.

Having stated their preference children were asked to give their reasons.

At this stage they referred to the features which they had discussed earlier and were able to comment on the features of the materials which they felt made them user-friendlier. They were asked to comment on both the stimuli and the questions.

A1 (1996 Cover etc & S1/Q4, 5) and A2 (1999 Cover etc. & S1/Q1-9)
19 of the 24 children preferred A2 (the 1999 cover, contents, introduction, explanatory text with captions and questions 1-9, section 1). Gender was not an issue. 4 of the 5 exceptions were teacher assessed at level 4 for reading, with 1 at level 5.

Reasons given for preferring 1999 were
- content was more interesting and informative and therefore more purposeful
- 1996's sundial instructions were considered difficult and boring since there was no purpose as the model was not likely to be made
- pictures provided more support, especially as there were captions under the illustrations, making the details clearer and more accessible
- glossary and introduction gave more supportive information
- layout was clearer and easier to follow, whereas the sundial information needed to be re-read
- multiple-choice questions were quicker and easier to complete
- questions helped in locating information e.g. Q9
- variety of question types, including short answers, made it easier, rather than simply having to write the answers in a 'straightforward' way

Two of the exceptions who chose the 1996 version had a personal preference for making models.

B1 (1996 S1/Q1,2,3) and B2 (1999 S3/Q1-7)
22 of the 24 children preferred the 1999 information and reference text with cartoons. Gender was not an issue. The two exceptions were teacher assessed at level 4 for reading.

Reasons given for preferring the 1999 version were
- colourful cartoons made it look fun and more suitable for 11 year olds, whereas A1's information about clocks was considered more 'serious', factual and plain
- familiarity of the nursery rhyme theme made it more accessible and therefore made the questions easier
- headings in question form in the stimulus helped to provide cues
- larger print and more pictures made it easier and quicker to read
- picture cues in the questions helped in locating information in the text

Of the 2 children who chose the 1996 version, one boy felt nursery rhymes were not appropriate for year 6 and one girl felt that the humour was not funny.

C1 (1996 S2/Q1-20) and C2 (1999 S2/Q1-9 & S4/Q1-2)
The comparison of sections C1and C2 was more complicated. 12 children preferred C1 (1996's narrative) in total, 9 preferred C2 (1999's poem) in total. 2 level 5 children preferred 1996 narrative questions but 1999 poetry text. One level 3 girl did not complete this section. Overall the pattern showed an even split between the story and the poem. Most level 5 children preferred the narrative despite the length of the text.

Reasons given for preferring 1996 were
- preference for stories, especially with interesting themes e.g. talking clock
- finding C2's poem difficult to understand, and therefore the questions difficult to answer
- difficulty with C2's poem questions because of the descriptive nature of the poem, which required greater explanation about the meaning and use of words

- poems with rhyme and rhythm are easier
- C1's narrative questions were related more clearly and easily to the text and to what happened in the story
- multiple choice questions are easier and can be completed more quickly

Reasons given for preferring 1999 were
- shorter text takes less time to read and there is less to remember (mainly level 3 comment)
- it is possible to scan the poem to locate information
- layout of questions is better with enough space for answers (mainly level 3 comment)
- shorter answers make it easier (level 3 comment)

Conclusions

- There is no doubt that the children interviewed preferred the 1999 version of the test overall.

- This was especially the case for the first two paired sections compared, where the overwhelming majority preferred the 1999 version. In both cases the 1999 version's stimuli were thought more interesting and various layout features (e.g. pictures and captions, glossary and introduction, headings in question form, large print, picture cues in questions) provided recognisable support. In addition the nursery rhyme theme's familiarity was attractive, as were design features such as the use of colourful cartoons. The use of a variety of types of question was also appreciated because it was thought to help define the kind of answers required more clearly and assisted in the location of information in the texts.

- However the picture was more complicated for the third paired section. Here the comparison was made more difficult because of the differing nature of the text and question types. The children were split. Abler children often recognised that although the 1996 version's narrative text was longer than the poem in the 1999 version, the questions accompanying the narrative were easily related to the text whilst the questions on the poem were more sophisticated. Familiarity with the narrative form was also a factor and the poem was seen as relatively difficult to understand. Some less able children welcomed the relative brevity of the poem and the attractive layout of the questions – although they may not have appreciated their difficulty.

- Gender did not seem to be a factor governing children's preferences.

- Having considered both stimuli and questions, the children clearly perceived the 1999 paper to be more accessible and user-friendly than the 1996 version. How might this have affected performance?

- The children in the sample did not complete the tests and therefore it is not possible to say definitively how their perceptions might have changed if they had done so, or how their preferences might have affected their performance.

- Our evidence suggests that they would have felt more positive about the 1999 test than the 1996 test and it might be argued that such positive feelings could help them. Certainly they were able to identify features of both the stimuli and questions (especially in the 1999 version) which could assist their performance. Such features might make it more likely that they could produce the responses required by the mark scheme.

- However the data from the experimental comparisons of the 1996 and 1999 test results (which agrees with the results obtained when they were used operationally), indicate that the 1999 test (especially the Reading element) proved to be more difficult and demanding. In 1999 average marks were relatively low – as were the cut-scores governing the award of levels. The children may not have fully recognised the 1999 test's difficulty, although there was some evidence that abler children could sometimes do so. Attractiveness is not the same thing as accessibility. Children might find questions attractive and interesting when it is in fact relatively difficult to win the marks at stake. But of course difficulty in itself is not the same as severity in standards, as cut-scores are lowered or raised, as appropriate, to compensate for variations in question difficulty. But should those setting cut-scores also allow for greater or lesser user-friendliness too? We will return to this later.

- We can really only guess at the impact on performance of the improvement in user-friendliness and accessibility perceived by the children in this study. But the children seemed to think that the 1999 version was the more interesting and attractive and identified features that had been introduced to help them negotiate the tasks they were set. Without such features the 1999 test might have proved even more difficult.

4.5 KS2 Mathematics

Questions were selected from Paper A and Paper B of the 1996 and 2000 versions of KS2 mathematics, so as to ensure that interviews focussed on meaningful and productive aspects of the tests for the purpose of comparison. Pairs of questions were selected according to content and level to ensure that all national curriculum attainment targets were represented and that complications arising from variations in question difficulty would be limited as far as possible. The aim was also to avoid the repetition of questions with the same features, as this would have led to repetition and could have been frustrating for interviewees. Fifteen pairs of questions were included as detailed in Table 4.5.1.

Table 4.5.1 KS2 Mathematics Test - Paired Questions

Question Focus	*LEVEL 3*		*LEVEL 4*		*LEVEL 5*	
	1996	*2000*	*1996*	*2000*	*1996*	*2000*
probability			A10	A14		
collecting, representing, interpreting data	A6	A7	B12	B10	A14	A17
understanding and using measures						
understanding and using properties of position and movement			A13	A22		
understanding and using properties of shape	A5	A9	B9	B13	A18	B20
solving numerical problems	A3	A6				
developing an understanding of place value and extending the number system			B13	B9		
understanding relationships between numbers, developing methods of computation	B1, B4, B5	B1, B2, B4	A12	A12	B20	B18

**B1, B4, B5 ('96) were matched with B1, B2, B4 ('00) as they were short number questions.*

A Cambridgeshire school allowed the Project access to Year 6 children towards the end of the summer term of the 1999/2000 school year. A sample of 12 children were 'selected', with an even balance of gender and teacher assessed levels in mathematics attainment, as shown in table 4.5.2 below. The children familiarised themselves with the questions, and were then interviewed according to a semi-structured interview schedule, designed to help them to discuss the salient features of the materials.

Table 4.5.2 Interviewing 'Sample'

	Mathematics: Teacher Assessed Levels			
	Level 3	*Level 4*	*Level 5*	
n boys	2	2	2	6
n girls	2	2	2	6
n total	4	4	4	12

Children raised a range of similarities and differences whilst describing the selected materials. The following features were mentioned in descending order of frequency:

General Features

- context — the 'story' within which the mathematics is set
- response types — e.g. written explanations, ticks and crosses, circling answers
- contextualisation - — presence or absence, type e.g. pictures, words
- method box — where mark is awarded for working out
- layout — space, coloured areas, familiarity of question format
- marks — marks available for each question
- example answer — where one answer is completed
- bold font — emphasising important points

Question Specific Features

Shapes 1996 QA5 / 2000 QA9

- difficult, less familiar shapes in 2000-A9, 'normal' shapes in 1996-A5

Nets 1996 QB9 / 2000 QB13

- more squares in diagrams in 1996-B9, fewer triangles in 2000-B13
- picture of the box in 1996-B9 is helpful not 'just a picture', it is a clue
- in 2000-B13 you have to imagine

Grids 1996 QA18 / 2000 QB20

- a grid makes it easier 'you can turn the page around'

Co-ordinates 1996 QA13 / 2000 QA22

- mirror line is confusing, 1996-A13 is just co-ordinates, 2000-A22 brings in reflection
- letters and numbers on corners in 1996-A13, letters or numbers in 2000-A22

Graphs 1996 QB12 / 2000 QB10

- different scales
- different number of bars
- spacing between bars in 2000-B10

Spinners 1996 QA10 / 2000 QA14

- more spinners in 1996-A10
- each number appears once in 1996-A10, more often in 2000-A14

Preferences

1996-QA6 v 2000-QA7

Preferences were evenly divided (6 v 6)

Reasons for preferring 1996

- money question
- table

Reasons for preferring 2000

- chart has less information to handle
- less writing - quicker

1996-QA5 v 2000-QA9

More (n 7) preferred 2000-A9 than 1996-A5 (n 5).

Reasons for preferring 1996
- familiar question type
- shape
- selection of answer
- less writing
- no mirror - makes it quicker

Reasons for preferring 2000
- mirrors make it fun
- letter for answers

1996-QA3 v 2000-QA6

A majority (n 8) preferred 1996-A3 to 2000-A6 (n 4).

Reasons for preferring 1996
- picture
- shorter - less working out

Reasons for preferring 2000
- money
- mark for working

1996-QB1/4/5 v 2000-QB1/2/4

More (n 7) preferred 2000-B1/2/4 to 1996-B1/4/5 (n 5).

Reasons for preferring 1996
- more familiar
- more space on page
- answers in boxes are quicker

Reasons for preferring 2000
- 'playing with numbers'
- lines and circles in answers
- looks quicker

1996-QA10 v 2000-QA14

The majority (n 9) preferred 2000-A14 to 1996-A10 (n 3).

Reasons for preferring 1996
- answer is given - need to explain

Reasons for preferring 2000
- less to write
- 2 spinners make it hard
- layout/ bold font

1996-QB12 v 2000-QB10

The majority (n 8) preferred 2000-B10 to 1996-B12 (n 4).

Reasons for preferring 1996
- graph and table give more information
- estimating is more difficult

Reasons for preferring 2000
- money
- mark for working out

- box for working - can check mistakes

1996-QA13 v 2000-QA22
Preferences were evenly split (6 v 6).

Reasons for preferring 1996
- mirror line complicated question
- confused - reflection/symmetry
- shapes
- bold font

Reasons for preferring 2000
- set out better - more compact
- quicker - fewer co-ordinates

1996-QB9 v 2000-QB13
Preferences were evenly split (6 v 6).

Reasons for preferring 1996
- more familiar shapes
- triangles are difficult
- picture shows what to look for
- pencil in wrong place in B13
- easy words

Reasons for preferring 2000
- answer on lines
- fewer triangles than squares - easier to trace
- better layout

1996-QB13 v 2000-QB9
More (n 7) preferred 2000-B9 than 1996-B13 (n 5).

Reasons for preferring 1996
- table helps
- francs are difficult
- box for working out
- pictures

Reasons for preferring 2000
- easy words
- less money

1996-QA12 v 2000-QA12
The majority (n 9) preferred 2000-A12 over 1996-A12 (n 3).

Reasons for preferring 1996
- answer is 'just a number'
- fewer words to read - quicker, less to keep in head
- looks less/ plainer
- no speech

Reasons for preferring 2000
- more clues - words/ pictures
- bold font

- like story (word) problems

1996-QA14 v 2000-QA17
Preferences are evenly split (6 v 6).

Reasons for preferring 1996
- picture of can helps
- fewer words
- working out box

Reasons for preferring 2000
- pictures are fun, less boring

1996-QA18 v 2000-QB20
Almost all (n 11) prefer 2000-B20 over 1996-A18 (n 1).

Reasons for preferring 2000
- easier shapes
- prefer ticks/crosses

1996-QB20 v 2000-QB18
The majority (n 9) preferred 1996-B20 over 2000-B18 (n 3).

Reasons for preferring 1996
- space for working out
- bold font
- pictures
- like a story - more interesting

Reasons for preferring 2000
- fewer words
- 'just' write an answer
- no need to show working

Of the questions sampled, children tended to prefer the 2000 questions overall. There were 13 pairs of questions and in 2 cases more children preferred the 1996 question, in 7 cases more preferred the 2000 question and in 4 cases the choices were evenly split.

Of the 12 children 5 boys and 5 girls chose more of the 2000 questions, while 1 boy and 1 girl preferred more of the 1996 questions. However, for 7 of the children the difference was only one pair of questions. Overall there were no obvious gender differences or differences relating to ability, as estimated via teacher assessment levels.

Views on test taking
When asked 'What do you think about children being asked to do the tests?', 7 of the 12 children responded positively, commenting that it was good to test knowledge and that the information was useful for their next school. 5 of the 12 children responded negatively commenting that they found test-taking ‘scary’, stressful, depressing or boring.

Asked about their feelings before their own 'live' KS2 tests, all of the children interviewed were anxious. Their comments included worries such as difficulty of the tests; results; revision; lack of free time; negative comments from other pupils. After the tests 5 of the 12 children felt that they had been easier than expected and 7 of the 12 commented on their feelings of relief that the tests were over.

Summary

- Some of the features which were frequently mentioned when similarities and differences were described, did not influence children's preferences. For instance the context (the story or theme) of questions was the most frequently mentioned descriptive feature, but it did not seem to play a strong role in determining children's preferences.

- The main features picked out by children were related to the ways in which the questions were asked and the range of response types. These were key issues for children, who preferred a range of question and response types.

- Questions involving less reading, writing and explanation were preferred, as were more familiar question types. Since children were clearly conscious of the pressure of time, they favoured questions which they felt they could complete more quickly. These factors influenced the less able children to a greater extent.

- Another key feature was the 'working out box' which was often described as the 'extra mark box'. They had obviously been instructed by teachers to pay particular attention to this type of question and the fact that they could get an 'extra mark even with the wrong answer', was appealing. Some children felt strongly that even if marks were not available for working out, they would still rather have the space for working as they were anxious about 'doodling around the page' and 'making a mess' on the test paper. In such cases they worked in their heads, with the potential for mistakes which could not be checked.

- Presentational features and layout were also important, spacing of questions, large font and bold font made material more appealing. This was especially the case for the less able children.

- There were mixed views on contextualisation, with some children preferring words and pictures cues, while others felt that such cues were distracting and took more time to interpret. Children were able to distinguish between cues which were integral to the questions and those which were not. The less able felt that in some cases there were too many words to read, whereas in others they needed the words to interpret the mathematics. They seemed to sense when the balance impeded rather than facilitated their work and when the information was extraneous.

- The questions compared were only a sample from the four test papers and the differences in the number chosen by each child from the 1996 and

2000 versions were often small. However the balance of the evidence suggests that the 2000 questions were more appealing overall.

- There were no obvious patterns of response relating to gender or teacher assessment levels.

- Although there was considerable anxiety among children before the tests, most responded positively about their purpose and many felt that they had been easier than expected. They had clearly been influenced by the 'scare-mongering' of other pupils, even when teachers tried to allay their fears. It could be helpful for year 6 pupils to have direct feedback from pupils with a balanced view, possibly focussing on comments such as those made by the children in the study who felt that the tests were useful and easier than expected. This could help to reduce the stress for test-takers.

4.6 KS2 Science

Test materials were 'sampled' from the 1996 and 2001 KS2 Science tests, so as to ensure that interviews with children were focused on particular parts of the tests. As far as possible, pairs of questions, one from each version of the test, were chosen within each Science attainment target (except AT1), at each level. The selected materials were in 'paired sections', each containing 'comparable' 1996 and 2001 test materials, as detailed in Table 1 below. The amount of material under consideration during interviews, overall and at any one time, was thus limited to keep demands on children realistic.

Table 1 KS2 Science Test – Paired Sections

Attainment Target	*1996*	*2001*	*Focus of Comparison*	*Level*
Sc2.Life Processes & Living things	*A1* Test B Q6 'Identifying Animals'	**A2** Test A Q6a 'Seaweeds'	● ***Question Layout*** ● Font ● Spacing ● Use of bold	4
	B1 Test A Q6 'Exercise'	**B2** Test B Q4a to d 'Circulatory System'	● ***Question type*** ● Related Diagram	4/5
	C1 Test A Q4a to c 'Food chains'	*C2* Test B Q8a & b 'Butterfly Garden'	● Layout ● Related Diagram ● Reading	3/4
Sc3.Materials & Their Properties	*D1* Test B Q1 'Materials'	*D2* Test B Q1 'Absorbent Materials'	● Question type ● Related Diagram ● Reading ● Use of bold	3
	E1 Test AQ7a,b,c 'Powders'	*E2* Test AQ9a to d 'Mixing Materials'	● Question type ● Related Diagram ● Reading ● Question Layout ● Context	5
Sc4.Physical Processes	*F1* Test AQ9 'Sounds'	*F2* Test BQ11a&b 'Sounds'	● Question type ● Diagrams ● Reading	4
	G1 Test AQ3a&b 'Light'	*G2* Test AQ4a to c 'Lamp'	● Diagram ● Spacing ● Layout ● Question type ● Context	4/5

A Leicestershire school allowed the Project access to Year 6 children during the last half of the summer term, after the operational 2001 Key Stage 2 tests had taken place. A sample of 12 children were selected, with an even balance of gender and teacher assessed levels in science attainment. Due to unforeseen circumstances, however, some children were unavailable on the scheduled day and were replaced, when possible, at short notice; the number of children of each gender and level who were interviewed is shown in Table 2. The children familiarised themselves with the material and were then interviewed, using a semi-structured interview schedule.

Table 2 Interviewing Sample

	Science: Teacher Assessed Levels			*Total*
	Level 3	*Level 4*	*Level 5*	
Boys (n)	2	1	2	5
Girls (n)	1	2	3	6
Total (n)	3	3	5	11

Children pointed to a variety of features in describing similarities and differences between paired questions.

General Features

The following general features were mentioned, in descending order of frequency:

- question type — various ways of answering
- content — theme: eg:- materials, the body, animals, liquids, etc
- pictures/diagrams — presence, style, purpose, size, and quantity.
- context — presence and degree of information provided.
- layout — spacing, positioning of text and diagrams.
- labels — presence on diagrams
- reading demand — quantity of text, long sentences.
- title — presence

Section Specific Features

A1 (1996) Classification: 'Identifying animals' / A2 (2001) Classification: 'Seaweeds'.

- key - in both, used to compare two things
- style of Key - presentation using arrows (A1), or boxes (A2)

B1 (1996) 'Exercise' / B2 (2001) 'Circulatory System'

- graph - in B1

E1 (1996) 'Powders' / E2 (2001) 'Mixing Materials'

- table - in E2, its style & purpose
- skills required - E2 referred to as *'like comprehension'*

Preferences

After identifying salient features, children were then asked to state their preferences between the paired sections and, subsequently, to explain the reasons for their choices. In giving these the children often referred to the features they had discussed earlier and were able to comment on features of the sections that they felt helped them to engage with the questions. In reporting on this below, reasons given are cited in order of their frequency, those cited by most children being presented first.

A 1996-BQ6 v 2001-AQ6a

Most children (n 8) preferred the 1996 classification exercise (identifying animals) to the 2001 version (classifying seaweed) (n 3), although less able children (at level 3) more often preferred the latter. Layout seemed the critical feature.

Reasons for preferring 1996

- layout - aesthetically appealing; better key
- content - 'I know about animals'
- skill demanded - 'you don't have to look at the words (ie labels on diagram) which is something off your mind'; 'I know how to do it'
- cues – more information given (text)

Reasons for preferring 2001

- layout - set out clearer, with larger font

- cues - less information
- pictures text with the pictures is less helpful
- content - prefer plants

B 1996-AQ6 v 2001-BQ4a to d

Most children (n 8) preferred the 2001 example (about exercise) to the 1996 example in this pairing (n 3) (about the circulatory system)

Reasons for preferring 1996
- pictures - 'explains what he's doing' (ie taking his pulse)
- more information given
- cues - 'it explains more what the answer should be'

Reasons for preferring 2001
- question type - 'just ticking boxes'; choices given, 'no explaining'

C 1996-AQ4a to c v 2001-BQ8a & b

Preferences were split fairly evenly between the 1996 (n 6) question about food chains and the 2001 question (butterfly garden) (n 5).

Reasons for preferring 1996
- pictures - 'I can get the answers easier'
- cues - more information given; 'it explains more'

Reasons for preferring 2001
- pictures - more information given; easier to follow
- layout better
- content - 'happier about cycles'
- cues - 'learn more'
- familiarity - with similar questions

D 1996-BQ1 v 2001-BQ1

The 2001 question (absorbent materials) was much less popular (n 2) than the 1996 question on materials (n 9). Those preferring it tended to be able children who did so because it posed a challenge as well as being quick to answer via the mcq format.

Reasons for preferring 1996
- question type - less writing; quicker, 'choice for boxes helps'; clues given
- demand - less thinking to do; less questions
- content - 'more confident on solids, gases and liquids'

Reasons for preferring 2001
- demand - 'makes you think'; 'it is more complicated'

E 1996-AQ7a, b, c v 2001-AQ9a to d

The 2001 version question (mixing materials) proved more popular (n 8) than the 1996 one (powders - n 3) in this pairing.

Reasons for preferring 1996
- demand - more understandable
- question type - like multiple choice; quicker

Reasons for preferring 2001
- pictures - helpful; shows what is happening
- layout - set out better; more space; information in table easy to follow; answer space clearer
- cues - 'it tells you more about what is in the bags'

F 1996-AQ9 v 2001-BQ11a & b

Preferences were divided (1996 n 5 to 2001 n 6) between these two questions on sounds.

Reasons for preferring 1996
- pictures - 'real-life pictures help you more'; 'you can see what is going on properly'
- demand - less questions
- question type - one word answers

Reasons for preferring 2001
- question type - choices given; 'like doing flowchart'
- layout - easier; looks better
- pictures - more pictures; labels on pictures
- cues - 'it tells you more about what they are doing' (text)

G 1996-AQ3a & b v 2001-AQ4a to c

Again preferences were evenly divided (1996 (light) n 5 to 2001 (lamp) n 6). But those liking the 2001 version were mostly the abler (level 5) children, who appreciated the quality of the diagrams and the assistance they provided.

Reasons for preferring 1996
- question type - like multiple choice; drawing arrows is easier
- demand - less questions
- pictures - 'helps you see what the question is asking better'

Reasons for preferring 2001
- pictures - labels help; easier to follow; explains what is happening; look better
- content - ' I know quite a lot about shadows and reflections'
- question type - variety of writing and tick boxes

Views on test taking

Children were asked about their views and feelings about test taking. The majority responded positively, recognising that the tests were a means of establishing the level they had reached. Ideas also centred around the test being a form of preparation for secondary school ('*they know what level you're at and can sort you into classes'; 'it helps so you can be ready for hard questions')*. Two children commented that the tests were one means through which they could receive a good education and '*get a decent job'.*

Asked how they had felt before their own 'live' KS2 tests, the words *'nervous'* or '*scared'* were used by nine of the eleven children. One child did not respond, and the remaining child commented that he was fine. The reasons given for feeling nervous mainly concerned their fear of forgetting answers, or of not doing well, and for one child, this then raised fears of being *'put in a group I didn't want to be in'* (in secondary school).

The great majority (n=10) commented that they either felt *'much better',* '*glad', or 'relieved'* after they had completed their own tests. For all but one of these children, this feeling was associated with knowing that the tests were over - the exception relating to having done well. The remaining child in the sample still felt nervous, due to anticipation of the test results - not yet received at the time of interviewing.

Summary

- Overall, the materials drawn from the two versions of the test proved, broadly, equally popular with the children. Sometimes preferences were evenly split and 1996 and 2001 both provided examples which proved relatively popular.

- Certain features were frequently mentioned during the description of similarities and differences between the two tests. These included: the various ways of answering the questions; the content or theme; the presence, style and quantity of related diagrams/pictures; the degree of background information accompanying the question. To a lesser degree, the layout, labelling, title and reading demand of the questions were also mentioned as salient features.

- The features described above were frequently referred to when the children stated the reasons for their preferences. But specific features of individual questions (e.g. the keys in section A and the table in section G) also explained the children's preferences.

- Overall, for nearly half of the fourteen questions sampled, the children's primary reason for choosing a particular question was the related picture or diagrams. This was largely because they were seen to be a helpful aid by providing more information or demonstrating how or what occurred in an experiment. The children also appeared to prefer diagrams which were labelled (again because this was an additional aid) and those which were '*real-life'*, as this helped them to '*see what is going on properly'*.

- For four of the remaining eight questions sampled, 'question type' was the primary reason for preference, namely multiple-choice questions where the children could '*just tick boxes'* and were given choices for their answers, so that the questions became much quicker to complete.

- There were clear similarities in children's reasons for their preferences across the range of questions sampled. The following were mentioned repeatedly: pictures/ related diagrams, question type, demand (e.g. less thinking involved), content (either for reasons of familiarity/confidence, or because the children simply liked the theme), information cues and layout.

- Preferences did not seem to relate to gender but abler children (at level 5) appeared to have distinctive preferences in relation to a small number of sections.

- The majority of the children responded positively about their experience of operational KS2 testing. Most recognised that the tests were a means of establishing their achievement level which could be used for setting purposes at secondary school.

- However, the majority of the eleven children reported being nervous when they took their own KS2 tests, largely for fear of not doing well or fear of forgetting answers. After their tests the majority of the children simply felt relieved that the tests were over. The levels of anxiety described would seem to suggest that the children have been influenced (by parents, peers or teachers) such that they viewed their KS2 tests as ‘high-stakes’ assessments.

4.7 KS3 English

The intention of this small-scale qualitative study was to provide some empirical evidence concerning children's reactions to materials selected from the 1996 and 2001 versions of Paper 1 of the KS3 English Tests, to investigate the user-friendliness of features of the tests and ways in which variations in features of the test materials may make them more or less accessible, so affecting achievement.

Pupil's views were probed as they considered selected test materials, chosen to highlight variations in features which may facilitate or impede performance. KS3 English Paper 1 tests reading and writing and targets levels 4 – 7. The test materials chosen were divided into 'paired sections' of 1996 and 2001 stimuli and questions (A to G, as shown in table 4.7.1 below) which were presented separately for pupils to compare.

Table 4.7.1 KS3 English Paper 1 - Paired Sections

1996	*2001*
A1 front page, including instructions	A2 front page, including instructions
B1 Section A text extract (account of Arctic travel)	B2 Section A text extract (newspaper report of total eclipse of sun)
C1 Section A question 1 (How does she build up ...)	C2 Section A question 2 (How does the writer try to show ....)
D1 Section A question 2 (Explain her mixed thoughts and feelings about this experience)	D2 Section A question 1 (In what ways do the headline & lines 1 to 13 provide an effective introduction ...)
E1 Section B text extract (advertisement for an Antarctic cruise)	E2 Section B text extract (eyewitness account of 1883 eruption on Krakatoa)
F1 Section C question 4b (Write about someone who is frightened or nervous but who tries to overcome these feelings)	F2 Section C question 4a (Write about an experience you think you will never forget)
G1 Section C question 4c (Write about whether you think people should take part in dangerous activities like these)	G2 Section C question 4b (Write a report for a newspaper about a natural disaster.....)

A Cambridgeshire school kindly arranged for the Project to interview 12 pupils, selected to provide an even balance of gender and ability (see table 4.7.2 below).

Table 4.7.2 KS3 English interview sample

	level 4/5	*level 5/6*	*level 6/7*	*Total*
boys	2	2	2	6
girls	2	2	2	6
all	4	4	4	12

Pupils were asked to familiarise themselves with the selected materials and were then asked, initially to describe their features. The following general and question specific features of the test materials were mentioned, in descending order of frequency:

General Features

- titles and instructions - clarity, time allocation, detail, prompts
- font - type, size, bold, clarity
- bullet points - organisation, structure, clarity
- layout - spacing of texts and questions, accessibility
- level of detail - information about texts and questions
- line numbers
- length of text extracts and questions

Question Specific Features

A1 / A2 1996 & 2001 front pages, including instructions

- having the information in boxes draws attention to it
- national curriculum levels are in the middle of the page in A1 which makes it more noticeable
- the word 'test' is used in A2

B1 / B2 1996 & 2001 Section A text extracts

- it is better if the author's name is given

C1 1996 Section A Q1 / C2 2001 Section A Q2

- line references are given so you know where to look

E1 / E2 1996 & 2001 Section B text extracts

- in the 1996 version the colour makes the page brighter
- pictures make it more interesting, instead of just writing. They give more ideas
- there are different 'kinds' of writing – 2001 is an account. 1996 is more interesting

F1 1996 Section C Q4b / F2 2001 Section C Q4a

- in the 2001 version it says what will be assessed, so you can write better.

Preferences

Interviewees were then asked which of each pair they would prefer, and, subsequently, why?

A1 / A2 1996 & 2001 front pages, including instructions

The majority (10 out of 12) preferred the 2001 version, mainly citing layout issues as the reason for doing so.

Reasons for preferring 1996

- better layout
- easier to read

Reasons for preferring 2001

- bolder, clearer writing
- better layout
- more information/ explanation
- brighter colours

- boxes in 1996 version not liked
- time allocation clearer

B1 / B2 1996 & 2001 Section A text extracts

The 2001 report of the reception to the recent eclipse of the sun by crowds on a Cornish beach was chosen by 10 out the 12 pupils interviewed, in preference to an account of an encounter with polar bears by a female Arctic traveller.

Reasons for preferring 1996
- better content

Reasons for preferring 2001
- bold title
- more explanation
- easy introduction
- layout
- smaller paragraphs
- interesting content
- shorter text

C1 1996 Section A Q1 / C2 2001 Section A Q2

Whilst more (n 7) preferred the 2001 question, almost as many (n 5) would prefer to attempt the 1996 question in this quite similar pair of questions, both of which asked how the writer had produced specified effects and provided a series of prompts to support responses.

Reasons for preferring 1996
- clearer
- 'refer to .....' prompt
- easier content

Reasons for preferring 2001
- more detailed instructions
- bold font
- clearer - the paper is whiter
- more to 'go on'
- line references help
- 'refer to' instruction is more helpful at beginning than end

D1 1996 Section A Q2 / D2 Section A Q1

The 2001 question in this pair was selected by 10 of the 12 pupils involved. This too provided a series of prompts to help children explain how the writer produced a particular effect. The 1996 question in this pair was more open-ended, requiring children to rehearse the writers' thoughts and feelings about the events described.

Reasons for preferring 1996
- more open-ended
- short question

Reasons for preferring 2001
- more explanation
- more prompts
- bullet points help

- layout is clearer
- time advice helps
- content - prefer this to the 1996 question in this pair, which is about thoughts and feelings

E1 / E2 1996 & 2001 Section B text extracts

The 1996 stimulus materials (an advertisement feature for an Antarctic cruise, including colour photographs) were much the more highly regarded, being preferred by 11 of the 12 interviewees to 2001's contemporary account of the eruption on Krakatoa in 1883.

Reasons for preferring 1996
- more interesting
- better layout
- pictures and colour are eye-catching
- pictures give ideas
- easier to read

Reasons for preferring 2001
- content - prefer an account to an advertisement
- less to remember

F1 1996 Section C Q4b / F2 2001 Section C Q4a

The 2001 question was by far the most popular of these two (attracting 11 out of the 12 interviewees), with the reasons given suggesting that the more extensive supporting prompts provided in the 2001 question (which also used Section B as a foundation) may have swayed most.

Reasons for preferring 1996
- prefer content

Reasons for preferring 2001
- more explanation
- tells you what you will be assessed on
- better spacing and layout
- gives you more to write about

G1 1996 Section C Q4c / 2001 Section C Q4b

More (9 out of 12) preferred the 2001 question (asking for imaginative writing) to the 1996 question which required some argument of the issues surrounding participation in dangerous sports. Again the reasons given suggest that the provision of prompts and explanation of what was required in 2001 appealed to many interviewees, although others preferred the content of the 1996 question.

Reasons for preferring 1996
- more interesting
- factual - not made up

Reasons for preferring 2001
- more open question
- hints about what to write
- clearer targets
- clearer layout
- tells you how to set it out

Views on test taking

Most interviewees recognised that tests provided information for GCSE groupings, although there was some confusion about how 'important' the tests were. One felt that these tests were not for her benefit, but were for the school's. There was an even split between those who felt nervous and those who did not before their own 'live' tests. Another felt confident as he had had a lot of practice tests. All felt relieved after their operational tests, although five had some worries about the difficulty of the tests or concerns that they had not done well enough.

Summary

- Overall, having considered most of the contents of Paper 1 for both the 1996 and 2001 versions of KS3 English, interviewees preferred the 2001 version. Preferences expressed did not appear to relate to gender or ability level.

- One clear exception to this was a 1996 stimulus text (an advertisement feature about a cruise to Antarctica) which included coloured photographs. The colour, layout and content of this were seen as appealing.

- Clarity and layout of instructions, texts and questions were seen as important features, with fonts, spacing and the use of bullet points often mentioned by interviewees.

- The pupils interviewed appreciated the provision of detailed instructions and prompts to support them when constructing their answers. The 2001 version contained more of these. They particularly appreciated knowing the assessment criteria, which were also more explicitly stated in the 2001 writing questions.

- Interviewees reported some (but not high levels of) anxiety before and after taking their own operational KS3 tests. This was true for both boys and girls and for all ability levels.

4.8 KS3 Mathematics

Pairs of questions were selected from Tier 4 to 6 (papers 1 and 2) and Tier 5 to 7 (papers 1 and 2), according to content and level, to ensure that all national curriculum attainment targets were represented and that the complicating effects of question difficulty on comparisons would be limited. By presenting the materials as paired questions attention was targeted more effectively and the amount of material under consideration at any one time was limited to keep demands on the children realistic. The aim was also to avoid the repetition of question types with the same features. In all, fifteen pairs of questions were included, as detailed in Table 4.8.1.

Table 4.8.1 Mathematics Test - Paired Questions

Question Focus	*Tier 4 - 6 Qs*	*Tier 5 - 7 Qs*	*Focus of comparison*	
			Tier 4 - 6	*Tier 5 - 7*
understanding place value and extending the number system	P1 96 Q4 P1 00 Q5	P1 96 Q14 P1 00 Q7	Layout Ways of answering	Ways of answering Layout
understanding and using relationships between numbers and developing methods of computation	P1 96 Q2 P1 00 Q9	P1 96 Q9 P1 00 Q11	Pictures Contextualisation	Contextualisation* Context* Ways of answering
solving numerical problems	P1 96 Q9 ** P1 00 Q10	P1 96 Q6 ** P1 00 Q5	Layout	Layout
understanding and using functional relationships	P1 96 Q8 ** P2 00 Q7	P1 96 Q5 ** P2 00 Q3	Pictures / Diagrams Ways of answering	Context Ways of answering Pictures
understanding and using equations and formulae	P1 96 Q14 ** P1 00 Q11	P1 96 Q11 ** P1 00 Q6	Contextualisation	Contextualisation Layout
understanding and using properties of shape	P1 96 Q6 ** P1 00 Q6	P1 96 Q3 ** P1 00 Q1	Ways of answering	Ways of answering Contextualisation Diagrams
understanding and using properties of position, movement and transformation	P1 96 Q11 P1 00 Q3		Layout Language difficulty Ways of answering	
understanding and using measures	P1 96 Q3 P2 00 Q1	P1 96 Q2 P2 00 Q6	Layout Explanatn	Context Layout
processing and interpreting data	P2 96 Q15 P1 00 Q8	P1 96 Q10 P2 00 Q8	Ways of answering	Layout Diagrams Contextualisation
estimating and calculating the probability of events	P1 96 Q1 P1 00 Q2	P2 96 Q14 P2 00 Q17	Pictures / diagrams Ways of answering	Layout Contextualisation

** 'Context' refers to a question's setting, e.g. chocolate or marbles. 'Contextualisation' is the nature/extent of the context.*
*** indicates where Tier 4-6 questions were the same as the Tier 5-7 questions*

A Cambridgeshire school allowed the Project access to Year 10 pupils during the first half of the autumn term 2000. Due to the Project's schedule, we interviewed Year 10 pupils at the beginning of the school year about their experiences of the national tests in the previous term. A sample of 16 pupils were 'selected', with an even balance according to gender and ability (in terms of teacher assessment levels), as shown in table 4.8.2 below. The pupils familiarised themselves with the questions, and were then interviewed according to a semi-structured interview schedule, designed to help them to discuss the salient features of the materials.

Table 4.8.2 Interviewing 'Sample'

	Mathematics: Teacher Assessed Levels				
	Level 4	Level 5	Level 6	Level 7	total
n boys	2	2	2	2	8
n girls	2	2	2	2	8
n total	4	4	4	4	16

Pupils commented on various features of the questions when asked to describe them. The following were mentioned, in descending order of frequency:

General Features

- layout — The most frequently mentioned feature was spacing. Girls often commented on the use of bold. Font was mentioned, as was the lighter print on the 2000 version.
- content — Domain of the question, for example, probability; what you had to do, for example, estimate; whether examples were given.
- reading demand — The difficulty of the actual words, including the mathematical language; the need to read questions carefully and thoroughly; more reading was more time consuming.
- pictures — Level 7 pupils commented that pictures *didn't* help, whilst level 4 and 5 pupils often suggested that they *did* help.
- response types — The process involved in answering a question; the need to show working; ways of answering, e.g. sentence completion, fill in boxes, draw. Comments about multiple-choice questions were almost always by boys.
- question demand — The question's difficulty; the amount of working out required or the number of question parts.
- context — The nature of the context; how the context can relate to real life. There were many comments about context being unnecessary or irrelevant which came almost exclusively from level 7 girls.
- table — The inclusion of a table: the vast majority of comments related to tables being helpful.

Question Specific Features

1996 P1Q14 / 2000 P1Q7
- equations, addition and multiplication

1996 P1Q9 / 2000 P1Q11
- long equations make it look hard

1996 P1Q10 / 2000 P2Q8
- 2000 P2Q8 easier to read / see what's going on
- 2000 P2Q8 scale not good on graph
- graphs better on squared paper

1996 P1Q4 / 2000 P1Q5
- 2000 P1Q5 chart looks complicated
- 2000 P1Q5 clear, so you can see results

1996 P1Q11 / 2000 P1Q3
- different patterns
- clear graph in Q3

1996 P1Q3 / 2000 P2Q1
- graph paper - squares are clearer (Q1), dots (Q3) make you nervous

1996 P1Q9 / 2000 P1Q10 or 1996 P1Q6 / 2000 P1Q5
- 2 marks for each
- no calculator allowed in Q6 / Q9
- Q9 easier

1996 P1Q8 / 2000 P2Q7 or 1996 P1Q5 / 2000 P2Q3
- marks for working in both
- speech bubbles in Q7

1996 P1Q14 / 2000 P1Q11 or 1996 P1Q11 / 2000 P1Q6
- pencils show you where to write your answer

Preferences
Pupils were asked which of the paired questions (in the tier appropriate to their ability level) they would prefer to do and having expressed a preference were asked why. Their answers are summarised below. Reasons for preferences are only recorded here when given by more than one pupil, as a large number of diverse reasons were given.

1 1996-P1Q4 v 2000-P1Q5

The majority (n 6) preferred 2000-P1Q5 to 1996-P1Q4 (n 2). Boys preferred the place value question, asking them to complete a table by writing their answer in a box as opposed to having a plain white space to answer in.

Reasons for preferring 1996
- simpler
- less time consuming

Reasons for preferring 2000
- easier
- better layout
- better defined question

2 1996-P1Q14 v 2000-P1Q7

All pupils (n 7) preferred the short answer, completion type responses of 2000-P1Q7. The question kept the same format for all parts of the question, including the pattern in the use of bold.

Reasons for preferring 2000
- looks simpler
- language easier to understand

3 1996-P1Q2 v 2000-P1Q9

The majority (n 5) preferred 2000-P1Q9 to 1996 -P1Q2 (n 3).

Reasons for preferring 1996
- better explained
- looks less threatening

Reasons for preferring 2000
- better layout / looks simpler
- clearer, based around table

4 1996-P1Q9 v 2000-P1Q11

Preferences were evenly (3 v 3) split, with little in the way of explanation forthcoming.

Reasons for preferring 1996
- clearer

5 1996-P1Q9/P1Q6 v 2000-P1Q10/P1Q5

More (n 9) preferred 2000-P1Q10/P1Q5 to 1996-P1Q9/P1Q6 (n 5).

Reasons for preferring 1996
- looks interesting

Reasons for preferring 2000
- looks easier/ easier to follow
- lots of space for working out

6 1996-P1Q8/P1Q5 v 2000-P2Q7/P2Q3

Again the 2000 question (2000-P2Q7/P2Q3) was more likely to be preferred (n 10) to 1996-P1Q8/P1Q5 (n 5). Boys preferred the context of matchstick huts and the associated diagrams to the context of counters in the '96 question.

Reasons for preferring 1996
- table easier to understand
- good explanation
- shorter/ less information

Reasons for preferring 2000
- pictures help you understand
- looks easier/ more friendly

7 1996-P1Q14/P1Q11 v 2000-P1Q11/P1Q6

The 2000 equation (P1Q11/P1Q6) was much likelier to be preferred (n 12) than the 1996 question (P1Q6/P1Q11) (n 3). Boys in particular preferred the question with no context, simple layout and lots of white space.

Reasons for preferring 1996
- gives you more help

Reasons for preferring 2000
- simpler layout
- clear, no need to remember it
- less daunting
- less reading

8 1996-P1Q6/P1Q3 v 2000-P1Q6/P1Q1

Almost everyone (15 v 1) preferred 2000-P1Q6/P1Q1, which they perceived as the more familiar question, involving measuring and drawing angles - with no context, to the contextualised version.

Reasons for preferring 2000
- less complicated
- choices for answer - right or wrong

- better layout
- clearer instructions
- more familiar

9 1996-P1Q11 v 2000-P1Q3

All interviewees (n 8) preferred 2000-P1Q3, testing properties of position and movement, which had tick boxes for answering, diagrams on squared paper, a lighter font and more white space on the page.

Reasons for preferring 2000
- looks easier / clearer
- less reading
- simpler to understand

10 1996-P1Q3 v 2000-P2Q1

Here the majority (n 6) preferred 1996-P1Q3 to 2000-P2Q2 (n 2).

Reasons for preferring 1996
- looks simpler
- clearer

Reasons for preferring 2000
- gives examples
- layout better

11 1996-P1Q2 v 2000-P2Q6

Most (n 6) preferred 2000-P2Q6 to 1996-P1Q2 (n 2).

Reasons for preferring 2000
- less time consuming/ less to do
- looks easier

12 1996-P2Q15 v 2000-P1Q8

The 1996 question (P2Q15) proved more popular (n 6) than 2000-P1Q8 (n 2).

Reasons for preferring 1996
- easier
- bar charts simpler
- tick boxes
- clearer layout

Reasons for preferring 2000
- story
- easier

13 1996-P1Q10 v 2000-P2Q8

All pupils (n 7) preferred the version with diagrams on squared paper, with less contextualisation, fewer instructions and more white space on the page.

Reasons for preferring 2000
- more visual
- easier to read
- graph helps understanding

14 1996-P1Q1 v 2000-P1Q2

Most (n 5) preferred 2000-P1Q2, especially the boys, who preferred the question with choices to tick. Those (n 3) choosing 1996-P1Q1, (which had pictures of people with speech bubbles) included both level 4 girls.

Reasons for preferring 1996

- prefer pictures and text
- looks easier

Reasons for preferring 2000

- more reading and explanation
- clearer
- tick boxes

15 1996-P2Q14 v 2000-P2Q17

Everyone (n 7) preferred 2000-P2Q17. This had less contextualisation, more white space and more of a pattern in the use of bold.

Reasons for preferring 2000

- visuals help understanding
- less complicated/ easier to read
- less time consuming

Of the questions sampled, the pupils tended on the whole to prefer the 2000 versions. There were 15 pairs of questions. Overall pupils preferred the 1996 version in only 2 cases; choices were evenly split in 1 case and they preferred the 2000 version in the other 12 cases. Of the 16 pupils, all but one chose more of the 2000 questions, while 7 of the 16 chose either one or none of the 1996 versions. The preference for 2000 questions was stronger for boys than for girls. When considering differences in preferences by level, the higher level pupils favoured the 2000 versions slightly more than the lower level pupils.

On 4 pairs of questions, *all* of the children chose the 2000 version and, on one, 15 out of 16 chose the 2000 version.

Views on test taking

When asked what they thought about pupils being asked to do the tests, nine of those interviewed produced positive reasons for completing national tests. These included revision, and for teachers and pupils to know their levels. They felt this might influence setting procedures. Some pupils seemed unsure of the real purpose and consequences of the tests. Around a third of the pupils felt the tests were for the benefit of the teachers and not for their benefit. Half of the pupils expressed negative views. These mostly related to the pressure they felt. They also commented on how learning stopped and their time was spent revising, which they felt was less beneficial.

Asked about their feelings when taking operational KS3 tests, 13 of the 16 children interviewed experienced negative feelings before completing the tests. Boys and girls at all of the levels interviewed reported feeling nervous and scared. This had stemmed from them being unsure as to how difficult the tests would be and not really knowing what to expect. Three of the 16 pupils

reported that they didn't feel pressured. After completing the tests more than half of the pupils felt relieved that they were over. Seven of the sample interviewed commented that they had not been as hard as they had expected and that they felt satisfied with their performance.

Summary

- The questions compared were only a small sample from the papers for the different tiers, but overall the pupils preferred the 2000 questions. On one pair of questions the preferences were evenly split, on two pairs more pupils favoured the 1996 version and on 12 pairs more preferred the 2000 versions.

- The main features of questions picked out by pupils were: the way the question was laid out on the page, particularly the use of bold and the amount of white space on the page; content; reading demand; use of pictures; ways of answering; question demand; and context.

- Features frequently mentioned when similarities and differences were described were often reflected in the pupils' reasons for their preferences. Links were sometimes indirect. For example, if a pupil had commented on the differences in question layout, s/he may have given the clarity of a question as a reason for their preference.

- Clarity and ease seem to be the most important features. Comments referred to the questions looking or being easier or simpler, with clearer or simpler layouts and clearer explanations. Tables, diagrams and pictures were often regarded as helpful. Questions that were less complicated and, in particular, less time consuming were favoured.

- Where there were strong preferences for the 2000 version, either by boys only or by the whole sample, there were references to similar constructs. These included preferences for questions with little or no contextualisation; plenty of white space on the page; a regular pattern in the use of bold; graphs and diagrams presented on squared paper rather than plain white paper.

- Most pupils saw valid reasons for completing the tests, although some seemed uncertain as to the tests' real purpose. The majority of the year 10 pupils described feeling nervous or scared before the tests. However, when the tests were over, they felt relieved; commenting that the tests had not been as bad as they had expected. It may be prudent to clarify the *purpose* of the tests with pupils and to feed back the reactions previous pupils had *after* completing the tests to allay new test takers' concerns.

4.9 KS3 Science

This qualitative investigation sought to provide evidence of pupils' reactions to the 1996 and 2001 KS3 Science test materials. Semi-structured interviews were designed to probe pupils' subjective experiences and perceptions and to investigate features of the tests which might facilitate or impede their performance. Interviews were based on questions 'selected' from the Tier 3 to 6 test (papers 1 and 2) and the Tier 5 to 7 test (papers 1 and 2) to provide a meaningful and productive focus. Pairs of questions (one from each version of the test) were selected according to content and level to ensure that all national curriculum attainment targets were represented and that variation in question difficulty would be minimised. By presenting paired questions attention was targeted effectively and the amount of material under consideration at any one time was limited, to keep demands on the children realistic. It also avoided the repetition of questions with the same features, which could have been frustrating for interviewees. 6 pairs of questions were included to provide a basis for comparisons (A to F) as detailed in Table 4.9.1 below.

Table 4.9.1 KS3 Science - Paired Questions

	Question Focus	*Paired questions*	*Focus of comparison*
A	Sc 2/4,5 Life processes and living things	1996 Paper 1 Q2 2001 Paper 1 Q1	style of diagrams, features of layout, applying given information vs. retrieval
B	Sc 2/2 Humans as organisms	1996 Paper 2 Q14 2001 Paper 1 Q8	layout, ways of answering
C	Sc 3/1 Classifying materials	1996 Paper 2 Q4 2001 Paper 1 Q5	context, layout, ways of answering
D	Sc 3/1,2 Classifying materials Changing materials	1996 Paper 1 Q13 2001 Paper 2 Q14	context, graphs, tables, diagrams, drawing, applying given information vs. retrieval
E	Sc 4/2 Forces and motion	1996 Paper 1 Q17 2001 Paper 1 Q14	context, ways of answering, length of text
F	Sc 4/2 Forces and motion	1996 Paper 1 Q8 2001 Paper 1 Q6	ways of answering, explanation, use of table

A Cambridgeshire school allowed the Project access to Year 9 pupils during the second half of the summer term 2001. A sample of 12 pupils were 'selected', so as to provide an even balance according to gender and ability (in terms of teacher assessment levels), as shown in table 4.9.2 below. The pupils familiarised themselves with the questions concerned, and were then interviewed, using a semi-structured interview schedule.

Table 4.9.2 Interviewing 'Sample'

	Science: Teacher Assessed Levels			
	Level 4	*Level 5*	*Level 6*	*total*
n boys	2	2	2	6
n girls	2	2	2	6
n total	4	4	4	12

General Features

Pupils raised a range of issues. Note that at this stage they were not asked to state preferences between the questions in each pair, just to describe the materials. The following features were mentioned, in descending frequency:

- ways of answering - for example, matching, multiple choice, cloze, give reasons, explanation, drawing, short answer
- diagrams / illustrations - many comments concerned how these helped to explain the question, especially how they made the questions clearer. Others concerned their presence, how they were labelled, or how easy they were to interpret.
- information - how much information was given, how it was displayed (for example, in a table which was preferred by many) and the complexity of the information provided
- content - the knowledge and/or skills tested by the question
- layout - the most frequently mentioned features being bold font and spacing
- context - familiarity, whether or not it is a 'real-life' context, and how the context can help you to understand the question

Question Specific Features

Pupil's comments could all be seen as general features of test materials, their comments being more 'generic' for this key stage/subject than in others. There were therefore no question-specific features to report.

Preferences

Pupils were asked which of the paired questions they would prefer to attempt, and, having done so, why? Their answers are summarised below. Reasons are reported where they were cited by more than one interviewee, the most popular being presented first.

A 1996 P1Q2 v 2001 P1Q1

A slight majority (n 7) preferred the 1996 question to 2001's (n 5) in this pair.

Reasons for preferring 1996 question
- better key
- more information given

Reasons for preferring 2001 question
- simpler layout

B 1996 P2Q14 v 2001 P1Q8

Preferences were evenly split between these two questions (n 6 each)

Reasons for preferring 1996 question
- bigger, bolder writing

Reasons for preferring 2001 question
- more spacing
- visually more appealing

C 1996 P2Q4 v 2001 P1Q5

The 2001 question proved the more popular of these two (by 9 to 3).

Reasons for preferring 2001 question
- looks clearer
- table provides information
- cloze answers

D 1996 P1Q13 v 2001 P2Q14

Opinions were divided here, with 5 preferring the 1996 version, against 7 who preferred the 2001 question.

Reasons for preferring 1996 question
- shorter
- less complicated

Reasons for preferring 2001 question
- more spaced out
- better diagrams
- clear table

E 1996 P1Q17 v 2001 P1Q14
The 1996 question here was chosen by the majority (n 9) over the 2001 question in the pair (n 3).

Reasons for preferring 1996 question
- multiple choice question
- less thinking needed to answer
- less to write

F 1996 P1Q8 v 2001 P1Q6
Preferences were again divided, with 5 selecting the 1996 question to 7 selecting the 2001 question in this pair.

Reasons for preferring 1996 question
- multiple choice questions
- less to write and explain

Reasons for preferring 2001 question
- clearer diagrams
- better layout
- helpful tables

Views on test taking
Pupils were asked what they thought about being asked to take KS3 tests and how they felt before and after taking them.

Seven of the twelve pupils interviewed recognised positive reasons for completing the national tests. These focused on testing their ability against recognised or national standards. A third expressed a negative attitude to testing due to the pressure they felt, with the more able being worried about doing well.

Half of the pupils interviewed reported feeling worried or nervous before taking their own KS3 tests. Others were less anxious and one (level 6) child reported feeling confident. Although such feelings were not on the whole strong, the less able pupils seemed the more worried, whilst the more able were less concerned. Three quarters of the pupils reported that after completing their operational tests they felt relieved or relaxed. Three pupils (all boys) said that the tests had no real impact on them.

Summary
- The questions compared were only a small sample of all those in the 1996 and 2001 tests, so that we would need to be cautious of reaching any

conclusions about overall preferences between the 1996 and 2001 versions of KS3 Science. But in the event, for the questions sampled, there was almost an even split between preferences for 1996 and 2001 questions. There were 6 pairs of questions. Pupils preferred the 1996 version in 2 cases, the choices were evenly split in 1 case and they preferred the 2001 version in the other 3 cases. Of the 12 pupils, 5 chose more of the 2001 questions, while 2 of the 12 selected more from 1996.

- Gender and ability level seemed unrelated to the preferences expressed.

- The main features of the questions preferred by pupils were: the method of answering; illustrations and diagrams - specifically how they helped to explain the question; tables - particularly how they help pupils to work things out; the content area of the curriculum tested; the amount of information given; and the layout of the question.

- Features frequently mentioned when similarities and differences were described were often echoed in the pupils' reasons for their preferences. Layout, in particular, seemed to be an especially important feature. Questions that involved multiple-choice answers, requiring less writing, were considered to involve less thinking and were favoured. Tables were also positively received and considered a helpful way of presenting information.

- Most pupils considered national testing a valid activity, recognising the need to check what they knew. There was an even split between pupils who felt nervous before their operational tests and those who did not. But when their tests were over, most felt relieved.

5 Evidence from LEA standardised testing programmes

Introduction

The second quantitative research strand within the project was an attempt to find and evaluate evidence about the relationships between results from standardised testing programmes in Local Education Authorities and national test results for the same children. We sought data from 1996 onwards - to correspond with the period covered by our experimental comparisons.

This approach considers standards comparable if, on average, children from successive cohorts with the same standardised test scores obtain the same national test results. Effectively this treats the standardised test as a reference test, which is used as a common yardstick to compare levels awarded in the national tests set each year. This approach can be criticised, as we acknowledged earlier, but it may provide useful supplementary evidence on issues the experimental comparisons cannot address, including concerns about identifying when shifts in national test standards might have taken place.

Using 'available data' from individual local authority testing programmes enables us to compares the results in different LEAs, some using the same standardised tests and others using different ones; introducing notions of replication which might help us decide how confident we can be in conclusions drawn from such data.

The evidence available

Why and how do LEA's use standardised test data?
Some Local Education Authorities mount programmes of standardised testing, in which schools administer the same test to all pupils at a given stage (or stages) in their education and the data are collected by the LEA (who may well pay for the tests and their marking). Pressure to devolve spending power to individual schools has perhaps worked against such developments in the last decade or so and many authorities have no such programmes.

LEAs employ standardised testing for various reasons. Often they were first introduced to help the Authority decide how best to allocate resources between schools, especially provision for the least able children. The standardised tests thus provide a measure of the calibre and range of children in one school as compared with another. More recently, LEAs have begun to use such information as part of ‘school improvement’ programmes. The data on each school’s ‘entry’ is used as a control variable in analyses which show how schools compare in ‘adding value’, in terms of subsequent test or examination results; so that they can see if their efforts are bearing more or less fruit than those of other schools.

Both these uses require only that children in a given cohort take the same test at the same time in a given year, but in practice such programmes do tend to

retain the same standardised test for several years, so that using these data for comparisons over time becomes possible. However different LEAs choose different standardised tests, depending on their purposes, preferences and the choice available when their programme began. They test at different points in children's educational careers too, choosing these to suit local convenience.

The data stored by LEAs varies enormously. Because standardised test data are used to provide management information to compare schools, many Authorities store information only at the level of their schools, aggregating information from individuals to obtain averages for each school. The pace of technological change also has an effect. Even where the same standardised tests have been administered for several years and pupil level data have been collected and stored in computer-based form, LEA's databases for any but the most recent years are sometimes no longer accessible. Hardware or software changes occur, or the selection of information retained is incompatible with that from more recent years.

Some LEAs also hold data concerning performance on national assessments. In some cases these too may be held at the individual pupil level, especially for recent years, when such data have been collected nationally by the Department for Education and Employment. But it is not always possible to link national test and standardised test databases, even where an LEA holds both types.

Locating suitable LEA standardised test data for our purpose
We wished to see how data from standardised tests relate to the results of national assessments in different years. Therefore our first task was to locate LEAs who held computerised databases for both types of measure for the same pupils, which were capable of being linked at the pupil level over a sequence of years.

No national agency holds information about LEA's testing programmes or databases and test publishers proved unwilling to identify their customers. There was thus no easy means of finding active LEAs. On the Project's behalf, QCA appealed to Inspectors and Advisors responsible for assessment to contact the Project if they held data which they thought might prove suitable. Several did so. The project's staff also used their own experience and contacts to contact LEAs they thought might be active in this field, and subsequently contacted those who were rumoured by others to be active. In all we made direct contact with twenty LEAs likely to have been active in standardised testing in recent years.

Only three held databases which had a ready (via a system of LEA pupil identifiers) capacity to link standardised test data and national assessments. We should perhaps note that at a national level there had been some reluctance to introduce unique pupil identification numbers, which would make it easier to track progress as pupils move through and between schools. Three more LEAs were able to provide data files on both types of assessment which could be linked by 'fuzzy matching' (via various combinations of

schools' and pupils' names, gender and date of birth). Fuzzy matching inevitably resulted in some data loss (as did absence at the times testing took place) but for our purpose this was not fatal, as it was only necessary to identify data for large groups of children spanning the full ability range. They did not need to be 'representative' as we were not primarily concerned to estimate or compare pupils' average performance on either type of assessment. Instead we only needed to see how one related to another in different years. Sequences of years for which data were available were also very limited. However it is important to recognise that all the LEAs contacted were very helpful and supplied data if they held any which was suitable[1]. The sparseness of the data available simply reflects the fact that it is rarely held, especially in an accessible form.

Table 5.1 outlines the data available from the six LEAs, listing the standardised tests, national assessments and year groups and annual cohorts involved and the total number of pupils for whom matched data were available for all the necessary variables. The percentages against each year (where data were available) show how many of the children in the data files supplied by the LEA could be included in our analyses, after checks to ensure that fully matched data for all relevant variables were present. The final column describes the basis for merging and matching data for different types of assessment etc. in each case. In all except LEA 4, the data provided included gender, enabling stratification by sex within our analyses.

Table 5.1 The standardised testing data available

	National Assessment & Standardised Test	*Total n*	*1996 %*	*1997 %*	*1998 %*	*1999 %*	*2000 %*	*merge variables*
LEA 1	KS1 En: Suffolk Reading Test	29896	95	96	98	-	100	Pupil ID
	KS2 En: Suffolk Reading Test	29926	93	88	92	-	93	Pupil ID
	KS2 Ma: NFER Maths 7-11	20788	-	88	92	-	93	Pupil ID
LEA 2	KS2 En: NFER-Nelson GR Test	22,985	90	89	89	-	-	Pupil ID*
LEA 3	KS2 En: NFER RC Test DE	4,772	-	34	33	-	-	School, Name, DoB*
LEA 4	KS1 En: NFER Primary Reading	52,950	99	100	100	100	-	Pupil ID*
LEA 5	KS2 En: Suffolk Reading Test	17,963	-	-	78	83	85	School, DoB, Gender
	KS2 Ma: NFER Maths 7-12	17,971	-	-	79	83	85	School, DoB, Gender
LEA 6	KS2 En: CAT Verbal	13,904	56	62	68	-	-	School, Surname, MoB, Gender
	KS2 Ma: CAT Quantitative	15,747	57	60	67	-	-	School, Surname, MoB, Gender

* data supplied by LEA already merged

LEA 1

In the case of LEA 1 the Suffolk Reading Test (SRT) was widely used at a series of testing points, including 6+ (when LEA 1 uses SRT's Level 1 test form) and 10+ (SRT Level 3), which were, respectively, close to the KS1 and KS2 testing points. NFER Mathematics 7-11 Series tests were also used in this LEA at 10+, but less widely. The LEA's unique pupil identifiers enabled a high level of matching, with data loss arising only where pupils were absent for one or other assessment or where gender data were missing. The Suffolk Reading Scale gauges what children are able to read; producing an age

[1] The Project would like to thank all the LEAs for their help, especially those who supplied us with data at short notice. We regret that we cannot identify them as we assured them we would preserve the anonymity of their data.

standardised score and a reading age. It addresses a much narrower achievement domain than those assessed by the KS1 or KS2 assessments. In contrast the NFER Mathematics 7-11 test is an up to date test which is in line with the national mathematics curriculum. It thus addressed much the same domain as the KS2 Mathematics tests and teacher assessments.

LEA 2

All schools in LEA 2 administered the NFER-Nelson Group Reading Test and the LEA's unique pupil identifiers enabled a high level of matching. Data loss was brought about only by missing data for one of the variables concerned in our analyses, as in LEA 1. The NFER-Nelson Group Reading Test uses sentence completion items to monitor reading progress by showing how pupils are performing compared to their peer group. Like the Suffolk Reading Scale, it has a relatively narrow curricular focus compared to the KS2 National Tests in English.

LEA 3

The NFER Reading Comprehension Test DE had been used in LEA 3 for many years, although data before 1997 proved inaccessible. This test too has a relatively narrow curricular focus; being primarily designed to monitor reading standards. The LEA were kind enough to match data from their standardised testing programme and their national assessment database before supplying the data to us, using children's names as the basis for this. Children wrote their own names on standardised tests, using various conventions for initials, first and second names and ordering, so the level of matching managed by the LEA was low - about 35% (and missing data subsequently took an additional toll). This notwithstanding, both cohorts in the data available to the project spanned the full ability range and were large enough for our analyses.

LEA 4

LEA 4 used unique pupil identifiers and their database was able to supply fully matched data for four successive cohorts. They employed the NFER Primary Reading Test in Year 2. This assesses children's ability to understand words and simple sentences and can be administered orally if required. This test too addressed a narrower curricular range than KS1 national assessments.

LEA 5

LEA 5's data from standardised testing and national assessments required matching via a combination of school codes, surnames and date of birth. A reasonably high level of matching was obtained (around 80%, after further attrition because of missing data). The Authority's testing programme on entry to secondary schools employed both the Suffolk Reading Test (Form 3) and NFER Mathematics 7-12. As the predecessor of the Mathematics test used in LEA 2, this test was less well matched to the recent developments in the national curriculum, but still provided a fairly broad assessment of achievement in mathematics.

LEA 6
The test used in LEA 6 was the Cognitive Abilities Test, from which both 'Verbal' and 'Quantitative' reasoning scores are reported. This test differed from the others in being expressly designed to assess potential for achievement, rather than the developed achievement measured by national assessments. Here too matching with national assessment records was required, using school, surname, month of birth and gender. Despite the fact that children had written their own names on the standardised tests the level of matching obtained was fairly good (85% +) within schools. However the data files available both had several (different) schools missing, which together with missing data led to the relatively low overall levels of matching reported. But here too the size and composition of the groups for each cohort were still entirely adequate for our primary purpose.

Data Analysis

Tables 5.2a, 5.3a and 5.4a respectively relate to KS1 Reading and Writing, KS2 English and KS2 Mathematics. Each presents a series of comparisons (within LEAs) of children's achievements in a given national assessment and a relevant standardised test.

In each case these show the numbers of boys and girls (where the data available distinguished between genders) for whom data were available for identified successive years, the means and standard deviations of their scores on the standardised test and their national assessment levels – by gender where available, together with various analyses relating to year on year changes in scores detailed in the illustrative example below. The correlations between standardised test scores and national assessments (also by gender) are also shown.

Key Stage 1 Reading and Writing

KS1 Reading Comprehension test levels
To explain the methods of data analysis employed, let us take the contents of the first comparison in table 5.2a (between KS1 Reading Comprehension test levels and Suffolk Reading Scale scores - administered in term 3 of Year 2, in LEA 1) as an illustration. All others follow the same approach.

These comparisons involved around 3,000 boys and similar numbers of girls from each of the 1997, 1998 and 2000[2] Year 2 cohorts assessed at the end of KS1 – 18,496 children in all (columns 2 and 3). Comparisons with other analyses for this LEA (also reported in table 5.2a) involving other national assessment measures (e.g. overall Reading and Writing levels) revealed that several hundred boys and girls were not entered for the Reading Comprehension (RC) test in each of these years, as would be expected.

[2] We were unable to obtain suitable data for the 1999 cohort.

The final column shows that the correlations between the standardised test (i.e. Suffolk Reading Scale) and this RC test were high and consistent over gender and cohorts – with the coefficients reported ranging between 0.70 and 0.73. These are quite high, especially when we consider that errors of measurement make perfect correlation impossible. For instance, if we assume that both the standardised and RC tests enjoy reliability coefficients as high as 0.9, then the maximum correlation value expected is 0.81. So the substantial correlation between these measures suggests that whilst they may not have been assessing exactly the same domain, there was probably considerable overlap.

In 1997 the boys' average Suffolk Reading Scale score was 103.3, with a standard deviation (sd, which estimates the spread of scores) of 12.1 (both in column 4). In 1998 the average score rose to 104.8 and by 2000 it had risen further still, to 106.5, with a standard deviation of 11.2. The boys' average KS1 RC test level was 2.37 in 1997 (the standard deviation then being 0.57 - both in column 6), rising to 2.48 (sd 0.48) in 1998, and rising again by 2000 to 2.52 (sd 0.39).

On average, achievement of children entered for the RC test (& note that this might involve selection effects which could vary over the years) on the Suffolk Reading Scale has thus risen over the period, as have their average KS1 RC test levels.

To help evaluate these year on year changes the differences between the Suffolk Reading Scale scores and KS1 RC test levels in subsequent years are each expressed as a percentage of their overall (across both genders and all cohorts available[3]) standard deviation. For instance the boys' average Suffolk Reading Scale age standardised score is 1.5 points more in 1998 than 1997 (104.8 - 103.3). This difference in standardised test scores between years amounts to 12.8% of a standard deviation (the ST YoY % sd change value reported in column 5). The rise in boys' KS1 RC test levels between the same cohorts was 0.11 levels (2.48 – 2.37), which amounted to 22.4% of a standard deviation (the NA YoY % sd change value reported in column 7). Thus whilst improvements have been registered in scores on both the standardised test and the national test in this instance (LEA 1 boys between 1997 and 1998) the changes in KS1 RC test levels were larger than those recorded for Suffolk Reading Scale scores when expressed in proportion to the spread of scores on each scale. But when the equivalent changes between 1998 and 2000 (1999 data being absent) are considered it is evident that this was reversed, with changes in boys' mean RC test levels being less than the increase in their standardised test scores (i.e. 8.2% v 14.6%, when both are expressed in proportion to their overall standard deviations).

The linear regression of RC test levels on Suffolk Reading Scale scores is used to estimate the average RC test level achieved by hypothetical 'average' children obtaining a standardised test score of 100. These too are reported in table 5.2a (as R est mean level @ ST=100 - in column 8). From these we can see that the linear model suggests that boys at this (same) point on the

[3] Not reported in table 5.2a.

standardised test score scale obtained average KS1 RC test levels of 2.25 in 1997. This estimated mean rose to 2.34 in 1998 but remained very similar (2.36) in 2000.

Table 5.2a Changes KS1 National Assessments and Standardised Test Scores by LEA

LEA 1: Suffolk Reading Test Level 1 (Y2,T3) with KS1 RC Test Levels

		n	Suffolk Reading mean / sd	ST YoY % sd change	KS1 RC L2 Test mean / sd	NA YoY % sd change	R est mean level @ ST=100	r reading/ KS1RC
Boys	*1997*	2913	103.3 / 12.1		2.37 / 0.57		2.25	0.72
	1998	3017	104.8 / 11.5	12.8	2.48 / 0.48	22.4	2.34	0.70
	1999	-	-	-	-	-	-	-
	2000	2954	106.5 / 11.2	14.6	2.52 / 0.39	8.2	2.36	0.72
Girls	*1997*	3192	104.6 / 11.2		2.46 / 0.54		2.30	0.71
	1998	3197	106.2 / 10.9	14.6	2.59 / 0.45	27.7	2.40	0.72
	1999	-	-	-	-	-	-	-
	2000	3223	107.7 / 10.5	13.7	2.60 / 0.37	2.1	2.40	0.73

ANCOVA: Year F ratio 87.3 (<0.001) / Sex F ratio 107.0 (<0.001) / Year * Sex F ratio 2.20 (ns)

LEA 1: Suffolk Reading Test Level 1 (Y2,T3) with KS1 Spelling Test Levels

		n	Suffolk Reading mean / sd	ST YoY % sd change	KS1 Spell. Test mean / sd	NA YoY % sd change	R est mean level @ ST=100	r reading/ KS1Spell
Boys	*1997*	3563	100.0 / 13.5		1.71 / 0.67		1.71	0.75
	1998	3591	101.7 / 12.9	13.1	1.81 / 0.69	14.7	1.75	0.74
	1999	-	-	-	-	-	-	-
	2000	3132	105.3 / 11.9	27.7	2.03 / 0.65	32.4	1.83	0.69
Girls	*1997*	3619	102.4 / 12.4		1.85 / 0.68		1.76	0.73
	1998	3567	104.2 / 12.1	15.0	1.97 / 0.68	17.6	1.80	0.73
	1999	-	-	-	-	-	-	-
	2000	3373	106.8 / 11.1	21.7	2.12 / 0.65	22.1	1.85	0.69

ANCOVA: Year F ratio 87.0 (<0.001) / Sex F ratio 56.0 (<0.001) / Year * Sex F ratio 1.27 (ns)

LEA 1: Suffolk Reading Test Level 1 (Y2,T3) with KS1 Reading Levels

		n	Suffolk Reading mean / sd	ST YoY % sd change	KS1 Reading mean / sd	NA YoY % sd change	R est mean level @ ST=100	r reading/ KS1Read
Boys	*1996*	3733	97.9 / 13.8		2.08 / 0.83		2.18	0.78
	1997	3716	99.2 / 13.8	9.5	2.13 / 0.77	6.4	2.16	0.78
	1998	3921	100.2 / 13.6	7.3	2.13 / 0.78	0.0	2.12	0.80
	1999	-	-	-	-	-	-	-
	2000	3720	102.5 / 13.1	16.8	2.22 / 0.74	11.5	2.12	0.77
Girls	*1996*	3673	100.8 / 12.8		2.30 / 0.73		2.22	0.78
	1997	3708	102.0 / 12.7	9.5	2.33 / 0.68	4.3	2.24	0.77
	1998	3753	103.2 / 12.7	9.5	2.35 / 0.71	2.9	2.21	0.80
	1999	-	-	-	-	-	-	-
	2000	3672	105.3 / 12.1	16.6	2.42 / 0.63	10.1	2.21	0.78

ANCOVA: Year F ratio 34.8 (<0.001) / Sex F ratio 245.6 (<0.001) / Year * Sex F ratio 0.75 (ns

LEA 1: Suffolk Reading Test Level 1 (Y2,T3) with KS1 Writing Levels

		n	Suffolk Reading mean / sd	ST YoY % sd change	KS1 Writing mean / sd	NA YoY % sd change	R est mean level @ ST=100	r reading/ KS1Writ
Boys	*1996*	3730	97.9 / 13.8		1.88 / 0.71		1.95	0.68
	1997	3721	99.2 / 13.8	9.5	1.91 / 0.69	4.2	1.94	0.67
	1998	3916	100.2 / 13.6	7.3	1.93 / 0.72	2.8	1.92	0.69
	1999	-	-	-	-	-	-	-
	2000	3719	102.5 / 13.1	16.8	2.03 / 0.71	14.1	1.94	0.67
Girls	*1996*	3670	100.8 / 12.8		2.10 / 0.59		2.00	0.68
	1997	3709	102.0 / 12.7	9.5	2.11 / 0.60	1.6	2.05	0.68
	1998	3754	103.2 / 12.7	9.5	2.15 / 0.63	6.6	2.04	0.70
	1999	-	-	-	-	-	-	-
	2000	3670	105.3 / 12.1	16.5	2.23 / 0.59	13.1	2.09	0.70

ANCOVA: Year F ratio 9.1 (<0.001) / Sex F ratio 480.6 (<0.001) / Year * Sex F ratio 1.47 (ns)

LEA 1: Suffolk Reading Test Level 1 (Y2,T3) with KS1 Reading Task Levels

		n	Suffolk Reading mean / sd	ST YoY % sd change	KS1 Read. Task mean / sd	NA YoY % sd change	R est mean Level @ ST=100	r reading/ KS1Read
Boys	*1996*	3732	97.9 / 13.8		2.08 / 0.83		2.18	0.78
	1997	2898	94.6 / 11.1	-25.7	1.89 / 0.70	-25.0	2.13	0.70
	1998	3023	95.6 / 11.2	7.8	1.87 / 0.71	-2.6	2.07	0.73
	1999	-	-	-	-	-	-	-
	2000	3722	102.5 / 13.1	53.7	2.22 / 0.74	46.1	2.12	0.77
Girls	*1996*	3673	100.8 / 12.8		2.30 / 0.73		2.27	0.78
	1997	2662	97.0 / 10.5	-31.2	2.06 / 0.63	-34.8	2.19	0.70
	1998	2473	97.2 / 10.6	1.6	2.01 / 0.66	-7.2	2.14	0.73
	1999	-	-	-	-	-	-	-
	2000	3674	105.3 / 12.1	66.5	2.42 / 0.63	59.4	2.21	0.78

ANCOVA: Year F ratio 71.7 (<0.001) / Sex F ratio 166.8 (<0.001) / Year * Sex F ratio 0.92 (ns)

LEA 1: Suffolk Reading Test Level 1 (Y2,T3) with KS1 Aggregated TA Levels

		n	Suffolk Reading mean / sd	ST YoY % sd change	KS1 Teacher Assessmnt mean / sd	NA YoY % sd change	R est mean level @ ST=100	r reading/ KS1TA
Boys	*1996*	3734	97.9 / 13.8		1.80 / 0.64		1.86	0.67
	1997	3724	99.2 / 13.8	9.5	1.84 / 0.64	6.3	1.86	0.68
	1998	3924	100.2 / 13.6	7.3	1.89 / 0.65	7.9	1.88	0.69
	1999	-	-	-	-	-	-	-
	2000	3630	103.0 / 12.9	20.5	2.01 / 0.58	19.0	1.92	0.69
Girls	*1996*	3673	100.8 / 12.8		1.96 / 0.58		1.94	0.66
	1997	3713	102.0 / 12.7	9.5	2.00 / 0.58	5.3	1.94	0.66
	1998	3758	103.1 / 12.7	8.7	2.06 / 0.62	8.0	1.95	0.69
	1999	-	-	-	-	-	-	-
	2000	3637	105.5 / 11.9	19.0	2.16 / 0.56	13.3	1.99	0.68

ANCOVA: Year F ratio 22.9 (<0.001) / Sex F ratio 194.5 (<0.001) / Year * Sex F ratio 0.75 (ns)

LEA 4: NFER Primary Reading Test Level 1 (Y2,T2) with KS1 Reading Levels

	n	NFER Pr. Reading mean / sd	ST YoY % sd change	KS1 Reading mean / sd	NA YoY % sd change	R est mean level @ ST=100	$r_{reading/KS1read}$
1996	13635	107.1 / 12.1		2.20 / 0.78		1.85	0.75
1997	13140	107.7 / 11.8	5.0	2.22 / 0.75	2.6	1.86	0.75
1998	13410	108.3 / 12.0	5.0	2.21 / 0.75	- 1.3	1.82	0.76
1999	12765	109.0 / 12.3	5.8	2.28 / 0.74	9.2	1.88	0.74

ANCOVA: Year F ratio 20.8 (<0.001)

LEA 4: NFER Primary Reading Test Level 1 (Y2,T2) with KS1 Writing Levels

	n	NFER Pr. Reading mean / sd	ST YoY % sd change	KS1 Writing mean / sd	NA YoY % sd change	R est mean level @ ST=100	$r_{reading/KS1writing}$
1996	13600	107.2 / 12.1		2.02 / 0.70		1.73	0.70
1997	13129	107.7 / 11.8	4.1	2.01 / 0.70	- 1.4	1.69	0.70
1998	13391	108.3 / 12.0	5.0	2.01 / 0.72	0.0	1.65	0.71
1999	12739	109.0 / 12.3	6.0	2.06 / 0.70	7.0	1.71	0.69

ANCOVA: Year F ratio 35.3 (<0.001)

Girls' scores and levels are marginally higher than boys, but apart from this quite predictable difference, comparisons between years reveals a very similar picture for them too. Girls' average Suffolk Reading Scale scores rose year on year, as did their RC test levels. Increases in RC test levels were greater than those in the standardised test between 1997 and 1998 (proportionate to their standard deviations) but less between 1998 and 2000 - just like the boys.

Comparing estimated mean RC test levels for children scoring 100 on the standardised test is akin to holding standardised test scores 'constant', despite the changes observed between cohorts. In this example the standardised test scores seem to imply that children taking the RC test are in some sense 'better' each year, so we were not surprised that their average RC test levels also rose. But the estimated mean RC test levels provide the means for comparisons relating to a constant point on the standardised test score scale (100), so allowing for the shift in the quality (as estimated by the standardised test) of those taking the test in different years. These comparisons suggest that children who were 'equivalent' in terms of reading achievement as measured by the Suffolk Reading Scale obtained slightly higher KS1 RC test levels in 2000 and 1998 than in 1997 - by about one tenth of a level.

The statistical significance of these differences in successive cohorts' national test scores has been evaluated by means of an analysis of covariance (ANCOVA). The ANCOVA controls for differences between the various groups' Suffolk Reading Scale scores and the relative numbers of boys and girls involved each year before considering the likelihood that the differences in national test scores between boys and girls and between the cohorts might have arisen by chance. ANCOVA results are reported at the foot of the set of comparisons and show in this case that differences between cohorts (the Year F ratio) were indeed statistically significant (i.e. the probability that they

might arise by chance was below 0.001 - less than one in thousand), as were those between boys and girls (the sex F ratio), as we should expect. The interaction term between these factors (the Year by Sex F ratio) was not significant (ns) in this case.

We have so far considered the relationship between LEA 1's data concerning KS1 Reading Comprehension test levels and a relevant standardised test – the Suffolk Reading Test. We will now look to this LEA's data concerning changes in other national assessments related to language, including overall KS1 Reading and Writing levels, to see if they follow the same pattern in relation to the standardised test scores. We must also investigate how far data from the second LEA providing KS1 data to this study either confirm the evidence from LEA 1, or contradict it.

KS1 Spelling test levels
The second set of comparisons in table 5.2a again relate to LEA 1. The table shows the numbers of children and means and standard deviations of Suffolk Reading Scale scores and KS1 Spelling test levels for the cohorts taking the KS1 Spelling test in 1997, 1998, and 2000. The rising pattern of Suffolk Reading Scale scores in this LEA is again evident and mean Spelling test levels also rise (for both genders) steadily throughout these years. The magnitude of the changes in standardised test scores and Spelling test levels in this case look quite similar when both are expressed as proportions of their standard deviations.

When the estimated mean levels for children with standardised test scores of 100 are compared, 'equivalent' children (both boys and girls) obtained estimated mean KS1 Spelling test levels about 0.04 of a level higher in 1998 than in 1997. Mean estimated levels had risen further still by 2000 producing a total increase in estimated mean KS1 Spelling levels of about a tenth of a level over the period 1997-2000. These differences between years were statistically significant.

KS1 Reading levels
In LEA 1: The third set of comparisons in table 2.3a provide data concerning overall Reading levels and Suffolk Reading Scale scores for the children taking these tests in 1996, 1997, 1998 and 2000. Aside from the few children disapplied from the national assessments, we are now considering full cohorts, including 29,896 children in total, so the slightly lower mean standardised test scores are to be expected. In light of this, the continued presence of a strong rising trend in mean Suffolk Reading Scale scores between 1996 and 2000 is of special interest, in that as an average for the LEA's population it suggests that the children's reading achievement (measured by repeated administration of the same test in each year) has improved. We will return to this important point later.

Correlations between the standardised test and Reading levels are consistently very strong (ranging from 0.77 to 0.8), suggesting that they are measuring quite similar traits. Mean Reading levels rose in 1997; remained much the same in 1998; and rose again in 2000. But when both are

expressed as proportions of their standard deviations it looks as though the standardised test scores were rising faster than Reading levels. The regression estimates of mean Reading levels for children with a Suffolk Reading Scale score of 100 points across this period seem to confirm this view. The estimated mean Reading levels fall slightly between 1996 and 2000; more so for boys (-0.06) than girls (-0.01). Equivalent children have received slightly worse results. After controlling for reading achievement, the (negative) differences in levels awarded to cohorts from different years were statistically significant.

Thus the rise in mean Reading levels observed may be misleading, given that our prime interest is in standards set within national assessments. Once we have controlled for reading achievement via the Suffolk Scale it seems that the data from this LEA suggests that KS1 Reading levels for 'equivalent' children may have fallen since 1996.

In LEA 4: LEA 4 also provided data relating standardised reading test scores (in this case the NFER Primary Reading Test) and KS1 Reading, although their data did not discriminate between boys and girls and spanned the period 1996 - 1999. Here too standardised test scores for reading were higher in each successive year; rising by about 5% of a standard deviation each year. KS1 Reading levels also rose across the same period (even though a small fall was recorded in 1998) but their improvement was (proportionate to standard deviations) lower than that of the NFER Primary Reading scores, just as was the case in LEA 1. However in this case, although the estimated mean levels for children at a standardised test score of 100 fell in 1998 they rose again in 1999, to a little above their 1996 value (+0.03 of a level). These variations between cohorts were also statistically significant after controlling for reading achievement.

We have here two independent cases, both involving very large numbers of schools and children, where (different) reputable standardised measures suggest substantial continued gains in achievement in reading by KS1 children since 1996. Confirmation adds credibility.

Table 5.2b Overview of changes in KS1 Reading & Writing and Standardised Test scores

YoY changes in regression estimates of mean KS1 Reading levels at standardised test scores of 100

LEA		*1996 to 1997*	*1997 to 1998*	*1998 to 1999*	*1998 to 2000*
1	Boys	-0.02	-0.04		0.00
	Girls	+0.02	-0.03		0.00
4	Both	+0.01	-0.04	+0.06	

YoY changes in regression estimates of mean KS1 Writing levels at standardised test scores of 100

LEA		*1996 to 1997*	*1997 to 1998*	*1998 to 1999*	*1998 to 2000*
1	Boys	-0.01	-0.02		+0.02
	Girls	+0.05	-0.01		+0.05
4	Both	-0.04	-0.04	+0.06	

Although there is some similarity between the two Authorities' data relating reading achievement to KS1 Reading levels, via different standardised tests, the regression analyses produced slightly different outcomes. Year on year changes in estimated mean Reading levels for children at a standardised test score of 100 in both LEAs are summarised in table 2.3b. This shows how the two sets of estimates are not necessarily incompatible.

The similarity between the year on year changes observed in the two authorities in 1997 and 1998 is remarkable. There is little change between 1996 and 1997 but in 1998 levels awarded fall by about a third of a level. The pendulum may have swung back in 1999, for which only LEA 4 provided data. But even if this were so, the only data available for 2000 - from LEA 1, suggest it may well have been counteracted yet again in 2000, taking standards back to their 1998 level.

KS1 Writing levels
In LEA 1: The relationship between Suffolk Reading Scale scores and KS1 Writing levels in LEA 1 was less consistent than for Reading. The correlations with Writing were lower (0.66 to 0.69), as we should expect, but were again similar between genders and cohorts. Mean KS1 Writing levels for both boys and girls rose, year on year, modestly in both 1997 and 1998 and more substantially by 2000 (especially boys). But whilst for boys, just as for Reading, the national assessments for Writing improved (proportionately) less than standardised test results, the same did not hold for girls. Between 1996 and 2000, estimated mean Writing levels for boys at scale point 100 on the standardised test fell marginally, but for girls they improved by 0.9 of a level.

ANCOVA results indicated that, after controlling for gender and variations between cohorts in standardised test scores, differences in Writing levels awarded to different cohorts were statistically significant.

In LEA 4: Average KS1 Writing levels achieved in LEA 4 fell slightly in 1997, remained steady through 1998 and rose in 1999. Considered against the improving standardised test results in this LEA already described (in relation to Reading levels), improvements in KS1 Writing levels failed to keep pace with the rate of improvement in NFER Primary Reading scores in 1997 and 1998. Moreover their improvement in 1999 may not have been enough to regain the ground lost in earlier years, as in Reading.

Estimated mean Writing levels at a standardised test score of 100 fell in 1997 and 1998 and although they recovered in 1999 they did not return to the 1996's peak. After controlling for the rise in reading achievement the differences in Writing levels were statistically significant in this LEA too.

Year on year shifts in estimated mean writing levels at a standardised test score of 100 are also summarised in table 2.3b, illustrating the similarities in the data from the two LEAs. By 1998 children probably obtained lower Writing levels than those with equivalent reading scores had received in 1996, although the pendulum may have swung back since then, leaving mean levels

awarded in 2000 close to those obtained by children of similar quality (as estimated by reading achievement) in 1996.

KS1 Reading Task levels
In LEA 1 (the only source) the proportion of children providing KS1 Reading task data appears to have fluctuated since 1996, when such data were available for the whole cohort. In 1997 and 1998 KS1 task levels were only available for about 75% and 78% of the cohorts respectively, but the 2000 data included task levels for practically all children.

Mean Suffolk Reading Scale scores reflect the varying selectivity involved, with average scores lower in 1997 and 1998 but exceeding the 1996 mean in 2000. Average KS1 task levels fluctuate similarly, again reflecting selection effects.

Comparisons on the level playing field provided by mean estimated KS1 task levels for children at a standardised test score of 100 are more meaningful. These suggest substantial declines in the mean KS1 task levels awarded in both 1997 (about -0.06 of a level) and 1998 (about -0.06 further) with a partial recovery (of about +0.06) between 1998 and 2000. This leaves the KS1 task levels awarded (to children with equivalent Suffolk Reading Scale scores) in 2000 about 0.06 of a level below 1996's peak.

KS1 Aggregated TA levels for Reading and Writing
The upward trend in mean Aggregated TA levels between 1996 and 2000 in LEA 1 (the only source of such data) resembles the rise in Reading Scale scores. But when year on year changes are expressed as proportions of standard deviations the improvement in TA levels seems the more conservative. Comparison of the mean estimated TA levels for children at a standardised test score of 100 in each cohort confirms this, with changes in mean levels here proving less dramatic than the differences in 'raw' levels. Once we allow for the improvements in the quality of successive cohorts suggested by the Reading Scale scores, Teacher Assessments in fact appear to have been largely stable between 1996 and 1998 and only rose by about 0.04 of a level between 1998 and 2000.

KS2 English

Table 5.3a presents the analyses like those described above at KS1 of relationships between achievement in KS2 English and relevant standardised tests. These data were provided by five LEAs - 1, 2, 3, 5 and 6. All five LEAs provided data concerning achievement on the KS2 English Test. One LEA was also able to make separate Reading and Writing Levels available and three provided additional data concerning Teacher Assessments.

Table 5.3a Changes KS2 English and Standardised Test Scores by LEA

LEA 1: Suffolk Reading Test Level 2 (Y6,T2) with KS2 English Test Levels

		n	*Suffolk Reading mean / sd*	*ST YoY % sd change*	*KS2 En. Test mean / sd*	*NA YoY % sd change*	*R est mean level @ ST=100*	*r reading/ KS2 En*
Boys	*1996*	3310	98.1 / 13.7		3.52 / 0.77		3.61	0.75
	1997	3179	100.6 / 12.4	19.5	3.78 / 0.65	35.6	3.76	0.69
	1998	3546	100.5 / 12.7	-0.8	3.72 / 0.68	-8.2	3.70	0.70
	1999	-	-	-	-	-	-	-
	2000	3466	101.0 / 12.3	3.9	3.98 / 0.72	35.6	3.94	0.75
Girls	*1996*	3312	98.9 / 12.3		3.82 / 0.74		3.87	0.74
	1997	3106	100.3 / 11.7	11.8	3.96 / 0.68	19.4	3.95	0.72
	1998	3419	100.7 / 11.9	3.4	4.01 / 0.69	6.9	3.99	0.71
	1999	-	-	-	-	-	-	-
	2000	3588	101.1 / 11.5	3.4	4.17 / 0.72	22.2	4.11	0.74

ANCOVA: Year F ratio 439.4 (<0.001) / Sex F ratio 345.2 (<0.001) / Year * Sex F ratio 4.7 (<0.001)

LEA 1: Suffolk Reading Test Level 2 (Y6,T2) with KS2 English TA Levels

		n	*Suffolk Reading mean / sd*	*ST YoY % sd change*	*KS2 English TA mean / sd*	*NA YoY % sd change*	*R est mean level @ ST=100*	*r reading/ KS2 EnTA*
Boys	*1996*	3508	97.3 / 14.2		3.41 / 0.90		3.54	0.77
	1997	3583	98.3 / 13.9	7.2	3.54 / 0.87	15.0	3.62	0.77
	1998	3872	99.1 / 13.7	5.8	3.60 / 0.85	6.9	3.64	0.76
	1999	-	-	-	-	-	-	-
	2000	3769	99.6 / 13.4	3.6	3.68 / 0.84	9.2	3.70	0.76
Girls	*1996*	3451	98.5 / 12.5		3.71 / 0.84		3.78	0.74
	1997	3316	99.3 / 12.4	6.5	3.81 / 0.83	12.2	3.85	0.76
	1998	3606	100.0 / 12.5	5.7	3.82 / 0.81	1.2	3.82	0.73
	1999	-	-	-	-	-	-	-
	2000	3786	100.5 / 12.0	4.0	3.94 / 0.78	14.7	3.95	0.73

ANCOVA: Year F ratio 88.6 (<0.001) / Sex F ratio 1090.8 (<0.001) / Year * Sex F ratio 4.1 (<0.01)

LEA 2: NFER-Nelson Group Reading Test 2nd Edn. 6-14[4] (Y6,T2) with KS2 English Test Levels

		n	*NFER-Nelson Reading mean / sd*	*ST YoY % sd change*	*KS2 En. Test mean / sd*	*NA YoY % sd change*	*R est mean level @ ST=100*	*r reading/ KS2En*
Boys	*1997*	3678	99.8 / 15.4		3.79 / 0.70		3.79	0.64
	1998	3880	103.3 / 15.0	23.4	3.78 / 0.69	- 1.4	3.67	0.69
	1999	3951	104.3 / 14.7	6.7	3.91 / 0.71	18.3	3.76	0.70
Girls	*1997*	3776	101.7 / 15.4		3.95 / 0.70		3.90	0.65
	1998	3789	104.8 / 14.6	20.7	4.05 / 0.71	14.1	3.89	0.70
	1999	3911	105.5 / 14.1	4.7	4.11 / 0.70	8.5	3.92	0.69

ANCOVA: Year F ratio 37.6 (<0.01) / Sex F ratio 593.4 (<0.001) / Year * Sex F ratio 27.9 (<0.01)

[4] A revised and re-standardised version of the test was introduced in 1998, which may introduce a discontinuity into these standardised test scores.

LEA 3: NFER Reading Comprehension Test DE (Y7,T1) with KS2 English Test Levels

	n	NFER RC DE mean / sd	ST YoY % sd change	KS2 En. Test mean / sd	NA YoY % sd change	R est mean level @ ST=100	$r_{nferRC/KS2En}$
Boys 1997	1028	97.0 / 12.0		3.71 / 0.67		3.83	0.73
1998	1293	97.3 / 12.0	2.5	3.75 / 0.69	5.7	3.86	0.68
Girls 1997	1137	98.8 / 11.4		3.93 / 0.68		3.98	0.72
1998	1314	99.3 / 11.8	4.2	4.04 / 0.71	15.7	4.07	0.71

ANCOVA: Year F ratio 16.4 (<0.001) / Sex F ratio 154.5 (<0.001) / Year * Sex F ratio 4.36 (<0.05)

LEA 5: Suffolk Reading Test Level 3 (Y7,T1) with KS2 English Test Levels

	n	Suffolk Reading mean / sd	ST YoY % sd change	KS2 En. Test mean / sd	NA YoY % sd change	R est mean level @ ST=100	$r_{reading/KS2En}$
Boys 1998	2811	106.0 / 14.2		3.84 / 0.72		3.62	0.71
1999	3350	105.8 / 13.9	- 1.4	3.99 / 0.70	20.8	3.79	0.70
2000	2821	106.5 / 13.9	4.3	4.13 / 0.72	19.4	3.89	0.72
Girls 1998	2793	105.9 / 13.0		4.11 / 0.70		3.88	0.71
1999	3376	106.1 / 13.0	1.5	4.18 / 0.69	10.0	3.94	0.71
2000	2752	106.2 / 13.1	0.8	4.30 / 0.70	17.1	4.06	0.72

ANCOVA: Year F ratio 294.3 (<0.001) / Sex F ratio 778.1 (<0.001) / Year * Sex F ratio 18.2 (<0.001)

LEA 5: Suffolk Reading Test Level 3 (Y7,T1) with KS2 English Reading Levels

	n	Suffolk Reading mean / sd	ST YoY % sd change	KS2 Reading mean / sd	NA YoY % sd change	R est mean level @ ST=100	$r_{reading/KS2read}$
Boys 1998	2811	106.0 / 14.2		3.99 / 0.73		3.79	0.66
1999	3349	105.8 / 13.9	- 1.4	4.19 / 0.70	27.8	3.99	0.68
2000	2821	106.5 / 13.9	5.0	4.40 / 0.67	29.2	4.19	0.68
Girls 1998	2793	105.9 / 13.0		4.23 / 0.70		4.02	0.67
1999	3376	106.1 / 13.0	1.5	4.30 / 0.69	10.2	4.08	0.68
2000	2752	106.2 / 13.1	0.8	4.47 / 0.65	24.7	4.26	0.67

ANCOVA: Year F ratio 500.6 (<0.001) / Sex F ratio 345.7 (<0.001) / Year * Sex F ratio 43.7 (<0.001)

LEA 5: Suffolk Reading Test Level 3 (Y7,T1) with KS2 English Writing Levels

	n	Suffolk Reading mean / sd	ST YoY % sd change	KS2 Writing mean / sd	NA YoY % sd change	R est mean level @ ST=100	$r_{reading/KS2writing}$
Boys 1998	2811	106.0 / 14.2		3.62 / 0.74		3.42	0.64
1999	3350	105.8 / 13.9	- 1.4	3.65 / 0.75	4.0	3.45	0.63
2000	2821	106.5 / 13.9	5.0	3.67 / 0.76	2.7	3.45	0.64
Girls 1998	2793	105.9 / 13.0		3.87 / 0.75		3.66	0.64
1999	3376	106.1 / 13.0	1.5	3.91 / 0.73	5.4	3.67	0.64
2000	2752	106.2 / 13.1	0.8	3.95 / 0.74	5.4	3.73	0.64

ANCOVA: Year F ratio 11.2 (<0.001) / Sex F ratio 939.5 (<0.001) / Year * Sex F ratio 2.3 (ns)

LEA 5: Suffolk Reading Test Level 3 (Y7,T1) with KS2 English TA Levels

		n	Suffolk Reading mean / sd	ST YoY % sd change	KS2 En. Test mean / sd	NA YoY % sd change	R est mean level @ ST=100	$r_{reading/KS2En\ TA}$
Boys	*1999*	3510	104.4 / 15.1		3.85 / 0.81		3.67	0.77
	2000	2950	105.1 / 15.0	4.7	3.94 / 0.81	11.1	3.73	0.75
Girls	*1999*	3430	105.4 / 13.7		4.08 / 0.77		3.85	0.76
	2000	2831	105.5 / 13.8	0.7	4.15 / 0.77	9.1	3.92	0.75

ANCOVA: Year F ratio .3 (45.0<0.001) / Sex F ratio 430.2 (<0.001) / Year * Sex F ratio 0.48 (ns)

LEA 6: Cognitive Abilities Test 'Verbal' (Y7,T1) with KS2 English Test Levels

		n	CAT Verbal mean / sd	ST YoY % sd change	KS2 En. Test mean / sd	NA YoY % sd change	R est mean level @ ST=100	$r_{CATver/KS2En}$
Boys	*1996*	2066	99.4 / 14.9		3.72 / 0.73		3.75	0.75
	1997	2566	100.3 / 13.1	6.6	3.94 / 0.65	31.4	3.93	0.69
	1998	3323	99.8 / 14.0	- 3.7	3.86 / 0.68	- 11.4	3.87	0.70
Girls	*1996*	2119	100.9 / 13.7		3.98 / 0.70		3.95	0.73
	1997	2355	100.8 / 12.6	-0.7	4.10 / 0.65	17.1	4.07	0.70
	1998	3375	101.4 / 13.4	4.4	4.18 / 0.68	11.4	4.13	0.71

ANCOVA: Year F ratio 155.0 (<0.01) / Sex F ratio 676.7 (<0.001) / Year * Sex F ratio 19.4 (0.001)

LEA 6: Cognitive Abilities Test 'Verbal' (Y7,T1) with KS2 English TA Levels

		n	CAT Verbal mean / sd	ST YoY % sd change	KS2 English TA mean / sd	NA YoY % sd change	R est mean level @ ST=100	$r_{CATVer/KS2EnTA}$
Boys	*1996*	2130	98.9 / 15.2		3.74 / 0.79		3.79	0.75
	1997	2664	99.7 / 13.4	5.7	3.87 / 0.72	17.3	3.89	0.70
	1998	3715	98.6 / 14.8	- 7.9	3.80 / 0.77	- 9.3	3.86	0.75
Girls	*1996*	2154	100.5 / 13.7		3.99 / 0.75		3.97	0.72
	1997	2406	100.4 / 12.7	- 0.7	4.08 / 0.71	12.0	4.06	0.71
	1998	3685	100.7 / 13.7	2.1	4.07 / 0.72	- 1.3	4.05	0.74

ANCOVA: Year F ratio 46.3 (<0.01) / Sex F ratio 517.8 (<0.001) / Year * Sex F ratio 0.17 (ns)

KS2 English Test Levels

In all five comparisons relating the KS2 English test (which was targeted at levels 3-5, although level 2 was also awarded to a small proportion of children) to (different) standardised tests, correlations between the national assessments and standardised test scores are fairly high (around 0.7) and consistent across LEAs, cohorts and gender.

In LEA 1: The data involve the Suffolk Reading Scale and span 1996 - 1998 and 2000[5]. In 1997 mean standardised test scores were 19.5% of a standard deviation higher than in 1996 for boys and 11.8% of sd higher for girls. Girls' Suffolk Reading Scale mean scores moved up a further 3.4% of sd in 1998, when boys' scores changed hardly at all (actually falling by 0.1 points, i.e. -

[5] Data for 1999 were not available in a suitable form.

0.8% sd). By 2000 mean Reading Scale scores have risen again, although modestly, by 3.9% of sd for boys and 3.4% of sd for girls.

However it seems likely that selection effects were at work. In 1996 95% of the children for whom KS2 TA levels were reported also took the KS2 English test. In 1997 test takers reduced to 91%, but recovered to 93% in 1998 and 2000. As it would be the weakest children who would be selected out - to be assessed via the Level 1-2 task rather than the Level 3-5 Test - the shifts in mean standardised scores here will in part reflect such changes in practice.

But the changes in standardised test scores can be compared with those of the KS2 Test results (for the same children). Compared to reading test scores, mean KS2 English test levels rose even more dramatically in 1997 - by 36.1% sd for boys and 19.4% sd for girls. Although the boys' mean English test levels fell back again somewhat in 1998 (when selection out was relaxed), they had improved again by 2000, as did those of girls. Improvements over the two years 1998 - 2000 matched the leap in levels between 1996 and 1997 - without any help from changing selection effects. The net effect was that in this LEA, both sexes recorded, proportionate to standard deviations, much greater improvements in mean KS2 English test levels than in Suffolk Reading Scale scores over the period 1996 to 2000.

The effects of this can best be seen by comparing the estimated mean test levels, for children at a standardised test score of 100, across cohorts. Such children received markedly higher test levels in 2000 than in 1996; the estimated mean level for boys being up 0.33 over this period, whilst that for girls rose by 0.24. The rises in estimated means came largely in 1997 and between 1998 and 2000. The changes in test levels observed were statistically significant after controlling for the improvement in standardised test scores.

In LEA 2: The data relate to 1997 - 1999. Mean scores on the standardised test (in this case the NFER-Nelson Group Reading Test) appear to have risen dramatically in 1998 compared to 1997, by 23.4% of sd for boys and 20.7 of sd for girls. However this should be treated with caution as a revised and re-standardised version of the test was introduced in 1998. This may have introduced a discontinuity and it is perhaps safest to disregard this first year on year comparison. But significant further improvements in mean NFER-Nelson Group Reading Test scores were recorded by both sexes in 1999 (6.7% of sd for boys and 4.7% of sd for girls). Year on year changes in mean KS2 English 3-5 test levels in 1998 varied by gender, with the boys' mean level falling very slightly whilst the girls' rose considerably. For both sexes, mean KS2 English test levels rose in 1999; proportionately, by more than their standardised test scores (i.e. by 18.3% of sd for boys and 8.5% of sd for girls).

When we compare the estimated mean KS2 English test levels for children with standardised scores of 100 each year in this LEA it appears that these fell in 1998 (especially for boys) but recovered in 1999, ending, overall, close to where they began.

In LEA 3: The data permit only one year on year comparison (1998 on 1997) on the basis of the NFER Reading Comprehension Test DE. Again however the standardised test scores improved modestly (by 2.5% of sd for boys and 4.2% of sd for girls). Mean KS2 English test scores also improved - by 5.7% and 15.7% of sd for boys and girls respectively. These improvements in KS2 English levels were again (proportionately) greater than the improvement in standardised test scores. Mean estimated KS2 English test levels for children at a standardised test score of 100 were slightly higher in 1998 than in 1997, with the change in mean test levels proving statistically significant after controlling for the improvements in standardised test scores.

In LEA 5: Year on year comparisons between 1998 and 2000 were available, using scores on the Suffolk Reading Scale as the yardstick for comparison. Between 1998 and 1999 the mean standardised test scores remain fairly constant. The mean for boys fell slightly, whilst that for girls rose to compensate. In 2000 Reading Scale scores again remained fairly stable, the boys mean recovering to just exceed the original, whilst girls' Reading Scale scores rose fractionally (by 0.1 scale points). Unfortunately the much lower number of children providing data in 1999 make these data unreliable as a means of comparing achievement over time in this LEA.

However, in contrast to the same children's standardised test scores, the mean KS2 English Test levels improved dramatically (by 20.8% of sd for boys and 10% of sd for girls) in 1999 and rose substantially again in 2000 (by 19.4% of sd for boys and 17.1% of sd for girls). The net effect was that mean estimated KS2 English test levels for children with standardised test scores of 100 rose substantially in both 1999 and 2000 - in total by 0.27 of a level for boys and 0.18 of a level for girls. After allowing for changes in standardised test scores, the differences in mean KS2 English levels between years were statistically significant.

This LEA also (uniquely) provided data allowing us to compare the separate KS2 English test levels for Reading and Writing. Against the background of relatively stable Reading Scale scores, the children's mean KS2 Reading levels rose substantially between 1998 and 2000 (by a total of 57% of a standard deviation for boys and 34.9% of sd for girls). When we control for changes in Suffolk Reading Scale scores by considering the estimated mean KS2 Reading levels of children with a standardised test score of 100, children obtained relatively high levels in 2000, with the linear model suggesting means of 4.19 and 4.26 in 2000 for boys and girls respectively, compared to 3.79 and 4.02 in 1998. (+0.4 of a level for boys and +0.24 of a level for girls). Mean KS2 Writing levels were comparatively stable over the same period; with estimated mean KS2 Writing levels for children with a Reading Scale score of 100 only rising by 0.03 of a level for boys and 0.07 of a level for girls.

In LEA 6: The data allowed comparisons over the period 1996-1998 on the basis of Cognitive Abilities Test (CAT) Verbal scores. This is an aptitude test and we might therefore expect scores to change relatively little over time, even if achievement were to rise. In 1997 the mean of CAT Verbal scores for

boys was up on 1996 (by 6.6% of sd) whereas the mean for girls was down (by -0.7 of sd). In 1998 this reversed, with the boys' mean falling compared to 1997's (by -3.7% of sd) and the girls' mean rising (by 4.4% of sd). These fluctuations no doubt combine random and sampling errors (the groups for different years included an element of selection by school, as described earlier) and are perhaps much as we might have anticipated for a test of this type. The boys' mean KS2 English Test level was substantially higher in 1997 than 1996 (by 31.4% of sd), as was the girls' (by 17.1% of sd). But whilst the boys' mean test level fell back (by – 11.4% of sd) in 1998, that of the girls rose again (by 11.4% of sd). Over the full period, the mean KS2 English levels for both sexes (expressed in proportion to their standard deviations) increased by more than their Verbal Aptitude scores, as we should expect given the expected stability of standardised test scores in this case. Estimated mean KS2 English test levels for children at a standardised test score of 100 for the 1997 cohort were higher (+0.18 and +0.12 levels for boys and girls respectively) than for the 1996 cohort. The girls' estimated mean test level rose again (+0.06) in 1998, although the boys' fell back (-0.06).

The year on year changes in estimated mean KS2 English test levels for children at a standardised test score of 100 for all LEAs are summarised in the first part of table 5.3b.

Table 5.3b Overview of changes in KS2 English levels and Standardised Test scores

YoY changes in regression estimates of mean KS2 En test levels at standardised test scores of 100

LEA		*1996 to 1997*	*1997 to 1998*	*1998 to 1999*	*1999 to 2000**
1	Boys	+0.15	-0.06	-	+0.24
	Girls	+0.08	+0.04	-	+0.12
2	Boys		-0.12	+0.09	
	Girls		-0.01	+0.03	
3	Boys		+0.03		
	Girls		+0.09		
5	Boys			+0.17	+0.10
	Girls			+0.06	+0.12
6	Boys	+0.18	-0.06		
	Girls	+0.12	+0.06		

* from 1998 to 2000 for LEA 1

YoY changes in regression estimates of mean KS2 English TA levels at standardised test scores of 100

LEA		*1996 to 1997*	*1997 to 1998*	*1998 to 1999*	*1999 to 2000**
1	Boys	+0.08	+0.02	-	+0.06
	Girls	+0.07	-0.03	-	+0.13
5	Boys				+0.06
	Girls				+0.07
6	Boys	+0.10	-0.03		
	Girls	+0.09	-0.01		

* from 1998 to 2000 for LEA 1

Data from LEA 1 suggests that for children of equivalent reading achievement (i.e. with a standardised test score of 100), estimated mean KS2 English test levels rose in 1997 in comparison to 1996's estimated mean level, especially for boys. LEA 6's data (in which we should perhaps place less confidence, because of fluctuations in the sample and because the logic for using an aptitude test to monitor changes in standards of achievement is weak)

produced very similar estimates of changes in levels achieved by equivalent children between the same years.

Data bearing on changes in estimated mean KS2 English test levels in 1998 are available from four LEAs, although we have reasons to doubt validity in two cases. In the two authorities where we can be most confident, (LEAs 1 & 3) girls fare better than boys, with the girls' mean estimated KS2 English test levels rising +0.04 of a level in LEA 1 (where boys' estimated mean levels fell by -0.06) and +0.09 (compared to boys' +0.03) of a level in LEA 3. LEA 2's 1998 data must be treated with caution, because a revised test was introduced then, but the data suggested that the boys' estimated mean test levels fell substantially whilst that for girls did so only marginally. LEA 6's data shows the boys' mean level fell whilst that for girls rose. In all, the data for this year show a mixed set of outcomes, suggesting that whilst, overall, test levels achieved by children of equivalent ability were similar to those achieved the previous year, there was evidence of a gender difference. On average, boys probably did less well in 1998, whilst girls did better.

Two sets of data relate to changes in 1999. Both show that improvements in mean KS2 English levels were proportionately greater than improvements in mean standardised test scores, especially for boys, who seem to have clawed back some ground lost on girls in 1998. Estimated mean KS2 English test levels for children of equivalent reading achievement suggest that boys gained 0.09 levels in LEA 2 compared to 0.17 in LEA 5, whilst girls gained only 0.03 levels and 0.06 levels in LEAs 2 and 5 respectively.

Two sets of data relate to 2000, although in one case the comparison is with 1998 (LEA 1 data for 1999 being unavailable). LEA 5's data recorded substantial improvements in KS2 English test levels between 1999 and 2000, for both boys and girls, much overtaking their small changes in Suffolk Reading Scale scores. Estimated mean levels for both boys and girls suggest that after allowing for changes in Reading achievement KS2 English test levels may have risen by about 0.1 of a level in 2000. LEA 1's data provided confirmation that 2000 KS2 English test levels appear to have risen by substantially more than improvements in standardised test scores might warrant over the period 1998 - 2000, suggesting that boys' test levels (after controlling for Reading Scale scores) have risen by 0.24 of a level and girls' by 0.12; only a little less than the cumulative changes in LEA 5.

The data from other LEAs suggest that LEA 1's data, which span the entire 1996 - 2000 period, may not be untypical. They suggest that children with equivalent reading scale scores have obtained better and better KS2 English test levels almost (as 1998 may have been an exception) every year since 1996, with an uplift of about a tenth of a level per year typical.

KS2 English Teacher Assessments
In LEA 1: The groups of children for whom KS2 English Teacher Assessments are available appear to be free of the selection effects affecting the KS2 English test. As such they provide a relatively safe basis for considering progress in reading achievement (as measured by standardised

test scores) in this LEA over time. The mean Suffolk Reading Scale scores for both boys and girls in successive cohorts between 1996 and 2000 rose consistently, although more so in 1997 and 1998 (cumulatively, 13% of sd for boys and 12.2% of sd for girls) than between 1998 and 2000 (3.6% of sd for boys and 4.0% of sd for girls).

Mean KS2 English TA levels also rose (by 21.9% of sd for boys and 13.4% of sd for girls) between 1996 and 1998 - which was (proportionate to standard deviations) by less than Reading Scale scores. But between 1998 and 2000 the increase in mean TA levels was 9.2% of sd for boys and 14.7% of sd for girls, proportionately greater than the rise observed in Reading Scale scores. Comparison of the estimated mean KS2 English TA levels for children with a standardised test score of 100, takes the improving Reading Scale scores into account. This shows that the average TA levels for children (both boys and girls) with equivalent Reading Scale scores rose fairly steadily between 1996 and 2000, although by markedly less than test levels increased in this LEA.

In LEA 5: The improvement in mean KS2 English TA levels (in proportion to standard deviations) between 1999 and 2000 was greater than that of Suffolk Reading Scale scores - for both sexes. Consequently, estimated mean KS2 English TA levels for children with a standardised test score of 100 reveal that children in this sense equivalent were likely to obtain higher levels in 2000 than in 1999; by 0.06 of a level in the case of boys and by 0.07 of a level in the case of girls.

In LEA 6: Standardised verbal aptitude test scores for the children providing data in this LEA in 1996 and 1998 were quite similar but although the 1997 cohort's mean aptitude test scores were higher, variations in the basis for sampling between cohorts make it dangerous to interpret this as representing any change in ability or achievement within this LEA as a whole. Mean KS2 English TA scores for both sexes for the children providing data in 1997 were also substantially higher than those for 1996 or 1998, as would be expected. The aptitude test scores do provide the basis for fair comparisons between years when we consider the estimated mean KS2 English TA levels for children with standardised test scores of 100. They suggest that on average the TA levels achieved by equivalent children were about a tenth of a level higher in 1997 than in 1996, but slightly lower in 1998 - a net effect much like the evidence from LEA 1.

The year on year changes in estimated mean KS2 English TA levels for these three LEAs are summarised in the second part of table 5.3b. The year on year changes are quite consistent across the three LEAs, with TA levels for equivalent children rising by almost a tenth of a level in 1997, falling back a little in 1998 and finishing slightly higher by 2000, in total perhaps about fifteen percent of a level above their starting point in 1996.

KS2 Mathematics

Table 5.4a displays the results of the analyses relating KS2 Mathematics national assessments to relevant standardised test data. Standardised testing in Mathematics is less common and only three LEAs were able to provide data (LEA 1, LEA 5 and LEA 6). Two of these were able to supply details of children's Teacher Assessments for KS2 Mathematics as well as their KS2 Mathematics test levels.

KS2 Mathematics test levels
In LEA 1: These data indicate that fewer schools participated in its mathematics standardised testing programme in 1997 and 1998 than took the standardised reading tests, although the numbers involved rose by 2000. The numbers of pupils involved in different years fluctuate so that it is likely that different schools are included, making it difficult to compare mathematics achievement between cohorts. The improvements in boys' and girls' mean NFER Maths 7-11 scores noted in 1998 (8.8% and 6.4% of sd respectively) and again in 2000 (over two years, 21.4% of sd for both boys and girls) might, in part or whole have arisen from such selection effects.

But this does not prevent our using the standardised test scores in mathematics to monitor changes in KS2 Mathematics test levels. Test levels fell, on average, in 1998, reversing the trend observed in the standardised test, which in this instance probably assessed much of the same curricular ground as the national tests. However, by 2000 KS2 Mathematics mean test levels had recovered to exceed their 1997 level.

When estimated mean KS2 Mathematics test levels for children with a standardised test score of 100 are compared, children in this sense 'equivalent' in mathematics achievement obtained lower KS2 Mathematics levels in 1998 than in 1997. Though these too recovered somewhat in 1999 they remained below those for 1997, by -0.04 and -0.07 of a level for boys and girls respectively. These disparities in mean levels (after controlling for the variations in standardised test scores) were statistically significant.

Table 5.4a Changes KS2 mathematics and Standardised Test Scores by LEA

LEA 1: NFER Maths 7-11 (Y6,T2) with KS2 Maths Test Levels

	n	*NFER Maths mean / sd*	*ST YoY % sd change*	*KS2 Maths Test mean / sd*	*NA YoY % sd change*	*R est mean level @ ST=100*	*r nferMal/ KS2 Ma*
Boys 1997	2337	101.3 / 13.3		3.89 / 0.70		3.83	0.78
1998	3084	102.5 / 13.4	8.8	3.82 / 0.75	-9.5	3.71	0.80
1999	-	-	-	-	-	-	-
2000	3294	105.4 / 13.7	21.4	4.02 / 0.75	27.0	3.79	0.81
Girls 1997	2202	100.9 / 12.0		3.83 / 0.68		3.79	0.76
1998	2897	101.7 / 12.5	6.4	3.73 / 0.71	-14.1	3.66	0.77
1999	-	-	-	-	-	-	-
2000	3386	104.4 / 12.9	21.4	3.92 / 0.72	26.8	3.72	0.80

ANCOVA: Year F ratio 0.6.0 (<0.001) / Sex F ratio 66.7 (<0.001) / Year * Sex F ratio 0.42 (ns)

LEA 1: NFER Maths 7-11 (Y6,T2) with KS2 Maths TA Levels

		n	NFER Maths mean / sd	ST YoY % sd change	KS2 Maths TA mean / sd	NA YoY % sd change	R est mean level @ ST=100	r nferMa/ KS2 MaTA
Boys	*1997*	2552	99.7 / 14.1		3.70 / 0.86		3.72	0.80
	1998	3306	101.4 / 14.1	11.8	3.81 / 0.83	13.1	3.74	0.80
	1999	-	-	-	-	-	-	-
	2000	3562	104.2 / 14.6	19.5	3.90 / 0.83	10.7	3.71	0.81
Girls	*1997*	2397	99.6 / 12.9		3.73 / 0.80		3.75	0.78
	1998	3086	100.6 / 13.3	7.5	3.77 / 0.79	5.1	3.74	0.77
	1999	-	-	-	-	-	-	-
	2000	3604	103.5 / 13.6	21.6	3.88 / 0.78	13.9	3.73	0.80

ANCOVA: Year F ratio 4.64 (=0.01) / Sex F ratio 3.0 (ns) / Year * Sex F ratio 2.2 (ns)

LEA 5: NFER Mathematics 7-12 , Test 11 (Y7,T1) with KS2 Maths Test Levels

		n	NFER Maths 11 mean / sd	ST YoY % sd change	KS2 Math Test mean / sd	NA YoY % sd change	R est mean level @ ST=100	r nferMa/ KS2Ma
Boys	*1998*	2878	107.2 / 15.3		4.00 / 0.77		3.71	0.81
	1999	3391	107.3 / 15.2	0.7	4.12 / 0.72	16.2	3.85	0.78
	2000	2843	108.8 / 15.6	9.8	4.19 / 0.70	9.5	3.87	0.79
Girls	*1998*	2750	106.6 / 14.3		3.93 / 0.73		3.67	0.78
	1999	3355	106.5 / 14.4	- 0.7	4.09 / 0.71	21.6	3.99	0.77
	2000	2754	108.0 / 14.3	10.5	4.11 / 0.70	2.8	3.81	0.78

ANCOVA: Year F ratio 161.1 (<0.001) / Sex F ratio 22.8 (<0.001) / Year * Sex F ratio 4.1 (<0.05)

LEA 6: Cognitive Abilities Test 'Quantitative' (Y7,T1) with KS2 Maths Test Levels

		n	CAT Quant. mean / sd	ST YoY % sd change	KS2 Math Test mean / sd	NA YoY % sd change	R est mean level @ ST=100	r CATQnt/ KS2Ma
Boys	*1996*	2084	100.4 / 13.8		3.89 / 0.80		3.84	0.76
	1997	2597	101.6 / 12.1	9.5	4.04 / 0.71	20.3	3.97	0.70
	1998	3256	101.7 / 13.1	0.8	3.95 / 0.74	- 12.2	3.85	0.75
Girls	*1996*	2126	99.4 / 12.5		3.81 / 0.76		3.87	0.75
	1997	2316	99.6 / 11.4	1.6	3.95 / 0.67	18.9	3.97	0.69
	1998	3368	101.1 / 12.4	11.9	3.90 / 0.73	- 6.8	3.88	0.72

ANCOVA: Year F ratio 82.1 (<0.01) / Sex F ratio 5.8 (<0.05) / Year * Sex F ratio 1.7 (ns)

LEA 6: Cognitive Abilities Test 'Quantitative' (Y7,T1) with KS2 Maths TA Levels

		n	CAT Quant. mean / sd	ST YoY % sd change	KS2 Maths TA mean / sd	NA YoY % sd change	R est mean level @ ST=100	r CATQnt/ KS2MaTA
Boys	*1996*	2109	100.0 / 13.8		3.86 / 0.80		3.69	0.73
	1997	2663	101.1 / 12.3	8.5	3.99 / 0.72	17.3	3.75	0.71
	1998	3680	100.8 / 13.9	- 2.3	3.94 / 0.78	- 6.7	3.74	0.73
Girls	*1996*	2157	99.1 / 12.5		3.87 / 0.75		3.71	0.70
	1997	2409	98.93 / 11.8	- 1.5	3.93 / 0.71	8.0	3.79	0.69
	1998	3640	100.4 / 13.0	11.5	3.95 / 0.75	2.7	3.75	0.74

ANCOVA: Year F ratio 29.2 (<0.01) / Sex F ratio 18.7 (<0.001) / Year * Sex F ratio 0.38 (ns)

In LEA 5: LEA 5's data relate to the period 1998 to 2000. Mean NFER Mathematics 7-12 scores were much the same in 1998 and 1999 in this LEA, but rose (by about 10% of a standard deviation) in 2000. Changes in mean KS2 Mathematics Test levels failed to match this pattern. KS2 Mathematics test levels were substantially higher in 1999 than in 1998, by 16.2% and 21.6% of a standard deviation for boys and girls respectively - when standardised test scores were fairly stable. In 2000, boys' KS2 Mathematics test levels rose again, by about the same margin (in proportion to their standard deviations) as their standardised test results. But the improvement in girls' mean test levels was more modest, despite the fact that their mean NFER Mathematics 7-12 scores had risen more than boys, so that in relative terms their mean national test levels in Mathematics fell back.

The net effect of these changes is best illustrated by the changes in estimated mean KS2 Mathematics test levels for children with standardised test scores of 100. Estimated mean KS2 Mathematics test levels for children of equivalent ability were markedly higher in 1999 than in 1998 (+0.14 levels for boys and +0.32 for girls). The boys' estimated mean rose again marginally in 2000 (by +0.02, whilst that for girls fell back (by 0.18), leaving them about 15% of a level higher than in 1998. The differences between years in levels awarded were statistically significant.

In LEA 6: The data revealed that the mean Cognitive Abilities Test (CAT) Quantitative scores of both boys and girls rose steadily between 1996 and 1998. This is perhaps surprising, given that this test is considered a measure of aptitude, and contrasts with the CAT Verbal scores for the same children, where no obvious trend seemed to be present. It is difficult to be sure that these shifts in mean scores are not merely sampling effects, as numbers fluctuated considerably between years. The number of children with valid test levels as a proportion of those with Teacher Assessments fell from 99% in 1996 to 97% in 1997 and 90% in 1998. If this reflected less able children being removed from testing, it might have helped explain the rising trend in standardised test scores. Yet mean KS2 Mathematics test levels rose in 1997 and fell in 1998.

The estimated mean KS2 Mathematics test levels of children at a standardised test score of 100 from each year take the variations in calibre of the children involved each year into account. These suggested that children of equivalent mathematical ability were awarded higher levels in 1997 than 1996 (by +0.13 of a level for boys and +0.1 for girls) before they fell back again in 1998.

The year on year changes in estimated mean KS2 Mathematics test levels for children with standardised test scores of 100 in different LEAs (which provide 'like for like' comparisons between cohorts) are brought together in table 5.4b.

Table 5.4b Overview of changes in KS2 Mathematics and Standardised Test scores

YoY changes in regression estimates of mean KS2 Ma test levels at standardised test scores of 100

LEA		*1996 to 1997*	*1997 to 1998*	*1998 to 1999*	*1999 to 2000**
1	Boys		-0.12	-	+0.08
	Girls		-0.13	-	+0.06
5	Boys			+0.14	+0.02
	Girls			+0.32	-0.18
6	Boys	+0.13	-0.12		
	Girls	+0.10	-0.09		

* from 1998 to 2000 for LEA 1

YoY changes in regression estimates of mean KS2 Ma TA levels at standardised test scores of 100

LEA		*1996 to 1997*	*1997 to 1998*	*1998 to 1999*	*1999 to 2000**
1	Boys		+0.02	-	-0.03
	Girls		-0.01	-	-0.01
6	Boys	+0.06	-0.01		
	Girls	+0.08	-0.04		

* from 1998 to 2000 for LEA 1

The standardised mathematical aptitude test data from LEA 6 provide the only data comparing 1996 and 1997. Children with equivalent mathematical aptitude appear to have been awarded higher levels in 1997 than in 1996 in this LEA. In 1998, LEA 6's data and that from LEA 1 (where the standardised test of achievement in mathematics used covers much of the KS2 mathematics curriculum) appear to provide mutual confirmation that equivalent children's KS2 Mathematics Test levels had fallen - by about the amount which might have been gained in 1997.

Data from both LEA 1 and LEA 5 are relevant to the two year period 1998 to 2000, and concur in supporting a rise in the KS2 Mathematics test levels then. Compromise between the observed values would suggest a net shift of about +0.1 of a level.

KS2 Mathematics Teacher Assessment levels
In LEA 1: The changes in KS2 Mathematics Teacher Assessment (TA) levels for the children from LEA 1 for whom such data were available in 1997 and 1998 show that when considered in proportion to their standard deviations, the higher mean levels in 1998 matched the differences in mean NFER Maths 7-11 scores in the two years. But although mean TA levels rose again between 1998 and 2000 the difference in standardised test scores between these two years appears (in proportion to standard deviations) even larger. Estimated mean KS2 Mathematics TA levels at a standardised score of 100 are consistent with this, with the linear model suggesting that children in this sense equivalent obtained marginally (-0.03 for boys and -0.01 for girls) lower TA levels in 2000 than in 1998. The differences in TA levels observed between 1996 and 2000 were statistically significant, after controlling for variations in mathematical achievement.

In LEA 6: These data relate to the period 1996 to 1998 and employ a quantitative aptitude test score as the yardstick to compare TA data from different years - between which sampling variations may have affected aptitude test scores. Comparison of the estimated mean KS2 Mathematics TA

levels for children with standardised test scores of 100 suggested that TAs were on average slightly higher (by 0.06 of a level for boys and 0.08 for girls) for equivalent children in 1997 than in 1996; but slightly lower in 1998 (by 0.01 for boys and 0.04 for girls).

A summary of the year on year changes in the estimated mean KS2 Mathematics TA levels for children with a standardised test score of 100 is provided in table 5.4b. Between 1997 (when LEA 6's data suggested that children might have had average TA levels higher than their equivalents had obtained in 1996) and 2000, TA levels for children of equivalent mathematical ability/achievement seem to have been very stable or even to have fallen slightly.

What can we conclude from these data?

Assumptions and logic

We need to consider what conclusions we can reach, on the basis of these data, about changes in the standards of achievement in schools (raising issues concerning children's levels of performance) and changes in standards set within national assessment (raising issues concerning the calibration of these measurement instruments). But before we can address either we must consider the assumptions and logic involved in using these data for such purposes.

Consider first questions about the test standards set in successive national assessments. In trying to shed light on this issue, the methodology we are employing depends upon the assumption that children who obtain given scores on the same standardised test should, on average, obtain similar results in national tests, irrespective of the year in which they take them.

For instance, if national assessment standards have been held constant over the period 1996 to 2000 and given that KS1 reading appeared to have improved so much in LEA 1, we might have expected KS1 Reading levels in this LEA to improve similarly. But we have already shown that whilst these rose they did not keep pace with improvements in Suffolk Reading Scale scores and that in consequence the levels achieved by children with the same standardised test score have fallen, year after year. Ergo, these data might be taken to suggest that more severe standards were applied in assigning KS1 Reading levels in 2000 than in 1996.

But does the chief assumption hold water? Should we expect that KS1/2 test achievement should improve (or decline) in keeping with standardised test scores? Maybe not. For instance, could teaching and learning effort be adjusted or refocussed over time, so that scores/levels on one might be expected to improve faster than (or even at the expense of) the other?

The national curriculum and its accompanying assessment regime were introduced as part of a drive to lever up learning standards. Teachers have

almost certainly increased their focus on meeting the targets set over recent years. If national assessment results were not improving in these circumstances, policy makers might well ask why not? Should we not have expected national test results (i.e. the distributions of levels awarded) to have improved over the last few years?

But if teachers' attention is directed towards these newer targets, is it not likely that they will have paid less attention to other areas, such as the domains addressed by standardised tests? Whilst schools are judged by their national assessment results, they have little incentive to coach children for tests used to assign special needs budgets or as input measures in value added analyses, where low scores might sometimes actually work to their institution's advantage. This quite convincing line of argument suggests that whilst we might have explained away faster increases by national test levels, the opposite pattern found in KS1 reading test data is strong evidence that standards for the award of KS1 Reading levels must have become more severe.

The general flaw in this reference test methodology, where changes on one measure are used as the basis for judging shifts in measurement standards on another, lies in the fact that it is usually possible to raise counter arguments. For instance, if better teaching and learning is raising attainment, then this is likely to make children better able to complete other correlated tasks, like standardised tests. If so we would therefore expect standardised test scores to rise too. Perhaps it is a matter of relative improvement, where we should expect standardised test scores to be relatively stable (i.e. to rise less quickly than curriculum targeted measures like national tests)? Is this universally true or are there circumstances where the improvements on curriculum focussed tests might be slower than on 'correlated' measures? Would anyone go so far as to argue that efforts directed to improving achievement in the complex achievement domains represented by the national curriculum and the Literacy Strategy could generate comparatively large improvements in standardised test scores? For instance, might standardised tests pose (increasingly) more straightforward problems than the tasks encountered in learning? Inevitably, resolving such issues becomes a matter of judgement.

After considering the plethora of issues raised above, our view is that with the advent of the national curriculum, national testing and the literacy and numeracy strategies, it is reasonably safe to believe that teachers will not have become more inclined to 'teach to' standardised tests since 1996. Neither does it seem likely that such tests have become better aligned to the school curriculum in recent years. We suspect that either little has changed in this respect or (even in LEAs which have continued to mount large-scale testing programmes) teachers might well have paid less attention to preparing children for the demands of standardised tests as the years have progressed.

National tests are a different matter. Teachers may well have become increasingly focussed on and/or better at teaching towards the national curriculum's learning goals, including - perhaps especially - those addressed

by statutory tests. As teachers have gained experience of the national tests they have probably become better at preparing children to take them, so that real improvements in learning and better targeted test preparation will be inextricably confounded. The effects of practice and preparation on test performance have been studied extensively. Bond (1993) reviewed this field and there would seem to be widespread agreement that limited gains in performance on aptitude and achievement tests can be produced by a modest amount of practice and coaching regarding the form of the particular test, with the abler and more naive students typically gaining most advantage. But the evidence elsewhere suggests that extensive practice in test-taking (beyond the one or two dry-runs that would be common in most schools in England) confers little extra advantage. The quantity of practice has probably been fairly constant over the time period we are concerned about, and whilst there may have been some improvements in coaching for national tests since their earlier years (when teachers might not have known what to expect) it seems unlikely that substantial gains from increasing familiarity would have continued beyond about 1997. By then most teachers have been familiar with the form of the tests. But curricular re-focussing and allied improvements in teaching are less likely to have reached an early plateau.

Consequently we would interpret improvements in standardised test scores as essentially conservative evidence of rising standards of achievement in schools. Where there is considerable overlap between the domains tested by a standardised test and curriculum focussed measure (e.g. Reading tests and KS1 Reading or KS2 Mathematics and standardised tests of achievement in mathematics) we might expect that any improvements in standardised test scores would if anything slightly under-estimate improvements in learning in schools. Where the match is less good (e.g. KS2 English and reading tests), under-estimation of improving standards of performance in schools again seems likely, but the size of any effect is difficult to judge.

So, if standardised test scores rose and national test levels fail to keep pace, we would feel relatively safe in concluding that test standards have become more severe. But if national test levels rise or remain stable against falls in standardised test scores, we cannot conclude that test standards have become more lenient, as this might be explained by increasing curricular emphasis on national test domains.

However much of the standardised testing evidence we have collected in fact falls into a less easily interpreted zone between these extremes. We have data from different authorities which suggest (with reservations in some cases concerning the stability of the samples, year on year) that standardised test scores are improving over time (indicating rising achievement) whilst national test results have improved too. It is the relative rates of change that are at issue. We could still be fairly confident where improvements in national test results seem to lag behind standardised test scores, implying severity in recent years. Our reservations are conservative with respect to conclusions in this direction. But if national test levels have improved faster than standardised test scores, with children achieving a given standardised test score obtaining higher national test levels in later years than those with

equivalent scores had previously enjoyed, this might be explained in part or whole by improvements in preparation or learning over the years in question. It becomes a question of degree: how much, if any, improvement in national test results can we justify?

These data cannot answer such questions and hence may not provide definitive comparisons of test standards over time. But fortunately they are only a part of the evidence gathered in this study, and the information they provide, though perhaps not conclusive, can still be used to see how far they support the results from the Project's experimental comparisons, with one methodology cross-validating the other. In this spirit, and in light of the assumptions and reasoning above, we have tried (below) to draw what conclusions we can from the standardised test data available to us.

Have children learned more over time?

This question concerns overall learning gains - or losses - in schools. Whilst not in itself about the Project's focus on national assessment standards, it is perhaps the most important question these data can help to answer. Do they provide any independent evidence that the national curriculum and the testing regime associated with it might have helped to raise standards of achievement in schools in England?

KS1 Reading
The standardised test scores (where pupils from successive cohorts have taken identical tests) provide strong evidence that in LEA 1, on average, boys' and girls' reading has improved, year on year, from 1996 to 2000. Moreover the shifts involved are far from trivial. The improvements in Suffolk Reading Scale scores recorded over the four years totalled 4.6 age standardised points for boys and 4.5 points for girls (about one third of a standard deviation in both cases). Given that the literature suggests that standardised test scores have changed very little over time it would be remarkable if this were widespread. We therefore looked to the data from LEA 4 (also sound evidence) with considerable interest to see if they provided confirmation. This they did, with NFER Reading Test scores improving by 1.9 scale points (about 16% of a standard deviation) in the three years between 1996 and 1999. The replication via evidence for an improvement of substantial magnitude from two LEAs using different standardised reading tests is convincing.

There seems no reason why these improved Reading Scale scores should be invalid. Indeed, as we have already argued, they may if anything under-estimate advances in the wider KS1 curriculum. It seems that KS1 children's reading achievement really has improved in recent years.

To appreciate the significance of these improvements in national assessment terms, imagine a cut-score (like those used to assign national curriculum levels) fixed somewhere near the mean score on a test. If average scores improved by 25% of a standard deviation (a compromise between the changes over time observed in the two sources of data available), about 10%

more children might reach such a cut-score. Where cut-scores are nearer the extremes of the score distribution, a smaller percentage would be involved.

KS2 Reading
There were several relevant sources of evidence here. LEA 1's Suffolk Reading Scale score data also provided a sound estimate of changes in the reading achievements of successive cohorts reaching the end of KS2. Mean standardised reading scores rose steadily for both sexes between 1996 and 2000, by 2.3 scale points for boys and 2.0 scale points for girls (16.6% and 16.2% of a standard deviation respectively). The very large improvements in standardised reading test scores recorded in LEA 2 in 1998 should perhaps be discounted, given the introduction of a new version of the test at this point. But within this LEA further improvements of 6.7% of a sd for boys and 4.7% of sd for girls were then recorded in 1999. These two data sets probably provide the soundest evidence of changes in achievement over time.

LEA 3's data for 1997-1998 involved greater levels of data loss whilst matching, which must weaken our confidence in their value as population estimates. However small increases in standardised scores were again recorded (2.5% of sd for boys and 4.2% of sd for girls) in the latter year. Similarly, even though LEA 5's data may also be affected by sampling variations between cohorts, they too suggest a modest advance in achievement on the Suffolk Reading Scale: 0.5 of a scale point for boys and 0.3 for girls (i.e. only 3.6% and 2.3% of sd respectively) over the two years between 1998 and 2000.

On balance, as at KS1, there is a consistent upward trend across these KS2 standardised reading test data from a variety of LEAs. This suggests that genuine gains in reading achievement have been made in recent years, although improvements in reading in KS2 may be smaller than those recorded in KS1.

To illustrate the potential magnitude of such effects on national test distributions, a shift in average standardised scale scores as low as about 10% of a standard deviation (a compromise between the observed changes) would lead to an additional 4% of the cohort reaching a level threshold set near the mean.

KS2 Mathematics
Two LEAs provided data from standardised achievement tests in mathematics which addressed much of the ground covered in the national curriculum. In neither case are the data ideal for the purpose of estimating changes in achievement over the LEA as a whole. In LEA 1 the largest dataset (relating standardised mathematics test scores to TA levels) still clearly falls short of the complete cohorts in some years, so that sampling effects may be confounded with year on year changes. However comparison of mean scores from each year suggests that standardised test scores improved substantially (by 4.5 and 3.9 standardised scale points for boys and girls respectively - i.e. around 30% of a standard deviation) over the four years 1996 to 2000. In LEA 5 the fluctuating numbers involved again suggest that sampling effects and

year on year changes might be confounded. Having acknowledged this risk, boys and girls standardised scale scores increased by 1.6 points and 1.4 points respectively (around 10% of a standard deviation) over two years between 1996 and 1998.

The evidence is therefore not conclusive, but it is very encouraging. The standardised tests in mathematics concerned address a substantial part of the KS2 mathematics curriculum, and it is unlikely that teachers will have increased their emphasis on this portion, or to have taken to teaching to these tests more actively. If anything the opposite seems more likely. Although the evidence comes from just two LEAs, their KS2 Mathematics national test results do not suggest they are untypical and it thus seems very possible that schools in England may also have made substantial gains in achievement in mathematics at KS2 recently.

To illustrate the potential impact on national test results of such changes, we should again note that an improvement of 25% of a standard deviation in mathematics test scores (again a compromise between the rates of improvement observed between 1996 and 2000) would result in about 10% of the children in the cohort moving across a hypothetical threshold fixed around the mean.

Have standards applied in national assessments changed over time?

The evidence above, showing that achievement levels in schools have risen, provides the background against which we must evaluate the stability of the standards set in national assessments over time - their calibration. If children are getting better we must expect that national test results will improve. The question is, how much improvement is justified?

KS1 Reading
Sound large-scale data from two LEAs using different standardised tests suggest that when improvements in standardised reading test scores have been controlled, children obtained slightly lower overall *KS1 Reading levels*, on average, in 1998 than in 1996 and 1997. Moreover, although Reading levels may have recovered in 1999, the evidence relating to 2000 suggests that they were then still at 1998's low point. There seems no reason to doubt this finding, given the overlap between the reference tests and KS1 Reading and the likelihood that any switch in teachers' efforts will have been directed towards the KS1 Reading test's domain, which would tend to produce the opposite outcome. It would thus seem safe to conclude that the standards applied in KS1 Reading assessments in 1996 have certainly been maintained and that it is possible that they have become more severe.

KS1 Reading assessments are quite complex, involving the Reading Task, and Reading Comprehension Tests at level 2 (which provides sub-divisions a, b & c within level 2) and level 3. LEA 1’s data allows us to explore matters a little further by considering the relationships between the Suffolk Reading Scale scores and achievement in these elements in the KS1 assessments. An

element of selection governs decisions about which children will be involved in each of them, many will not be assessed through them all and 'aggregation' is algorithmic.

The *KS1 Reading Task* must be used with all children judged to be working towards or in levels 1 and 2. In 1996 a Task level was recorded for virtually all children in LEA 1. Task levels were provided for fewer children in 1997 and 1998 but by 2000 they were again provided for virtually all children in this LEA. This may not be unrelated to the introduction of the Reading Comprehension Test, which became compulsory from 1997. It is clear that the task levels awarded to children with equivalent Suffolk Reading Scale scores fell in both 1997 and 1998. Again there is no reason to doubt the obvious implications. But why should this be? Might teachers conducting the Reading Task have become marginally less willing to give some children the benefit of any doubt? Could the introduction of the Reading Comprehension test as a second measure have had an effect, perhaps by making teachers more conservative? A partial recovery had been staged by 2000, but children were even then still receiving task levels more like those awarded to children of equivalent (Reading Scale) ability in 1997 than in 1996. KS1 Reading task assessments thus seem to have become marginally more severe over time.

In 1997 the *KS1 Reading Comprehension Test for level 2* ceased to be optional, as it had been in 1995 and 1996. It was required for all children who achieved level 2 on the Reading Task. LEA 1's data revealed that in 1997 about 78% of boys and 86% of girls in LEA 1 took the RC test. Slightly fewer used the 1998 version (77% of boys and 85% of girls) but more took the 2000 test (79% boys and 88% girls). Those taking the Test were, again by definition, likely to be relatively able, as their mean Suffolk Reading Scale scores confirmed. After controlling for reading achievement on the standardised test, it would appear that the levels awarded via Level 2 Reading Comprehension tests in 1998 and 2000 may have been a little higher than those awarded to equivalent children via the 1997 version - by perhaps about a tenth of a level. But is this not about what might be expected if teachers have re-directed some of their efforts to inculcating the skills required to perform well in these newly introduced high stakes tests? We would expect to see such improvements as teachers adapt in the early years of a new form of assessment, so these data may well be consistent with the test standards having been maintained. Note too that this does not conflict with the evidence that overall, KS1 Reading assessment has become more severe, as the system for allocating children to the elements used in the assessment of Reading at KS1 and for combining their outcomes to give a final Reading level allows alternative routes to the same result.

KS1 Writing

The evidence from the two LEA's data concerning *KS1 Writing* is remarkably consistent. They suggest that average KS1 Writing levels (for children of equivalent reading achievement in each year) fell in 1997 and again in 1998, and that a partial recovery in 1999/2000 failed to recover all the ground lost. This appears to have left KS1 Writing assessments relatively severe compared to those in 1996. Is this conclusion safe? Correlations between

KS1 Writing and standardised reading tests are lower than those with KS1 Reading, just as should be expected when the logical basis for association is so much weaker. But they are still relatively high (around 0.7) and very consistent across cohorts, genders and LEAs. Empirically, the association between the traits tested in both measures is quite strong. Given the outcome (i.e. implied severity), standardised reading scale scores would anyway seem likely to provide a conservative monitor, so there is no obvious reason to question the validity of this evidence. The form of the KS1 Writing assessments has been relatively stable over the period with which we are concerned and the criteria for the award of levels have remained unchanged. This might seem to guard against drift in standards, especially towards severity. But there are constant efforts to improve test materials etc. One potential disruptive factor was the introduction (in 1996) of the subdivision of level 2; into 2a, 2b and 2c. Teachers who might have awarded level 3 to abler children may have become progressively more likely to substitute 2a once this option became available. Improvements in the exemplification material available to teachers may have contributed to this trend. In 1997 and 1998 the exemplification provided showed work which quite definitely met the criteria for level 3. By 1999 it was considered necessary to include a borderline illustration, to show the minimum quality required at this level. Has concern that teachers might be too lenient been overdone? Might such factors have helped produce the pattern we observe, which seems to justify concern that KS1 Writing standards too might have become more severe since 1996?

The *KS1 Spelling Test* was administered to children whose Teacher Assessment in writing reached level 2 or who achieved this level (or higher) in the Writing Task. Entry of children judged to be at level 1 was optional. LEA 1's data record that the proportion of children taking this test has fallen. In 1997 98% of boys and 97% of girls were involved, but this dropped to 92% of boys and 95% of girls in 1998, and to 86% of boys and 92% of girls by 2000. After controlling for the effects of rising Suffolk Reading Scale scores over these years (especially strong here, reflecting increasing selectivity), it appeared that by 2000 mean KS1 Spelling levels were higher about 0.1 of a level higher than those obtained by children with equivalent Reading Scale scores in 1997. Either the spelling tests were becoming more lenient or children were getting better at spelling. As we have already argued that it is likely that the introduction of such tests would encourage teachers to focus on the knowledge and skills required to succeed on them, there seems no grounds to challenge the validity of these relatively modest improvements over the first few years of these spelling tests. But might continued improvements at this rate be harder to justify, given that we have probably already seen most of the benefits from switching attention and effort in this direction?

KS2 English Test

As at KS1 the null hypothesis assumes that KS2 test results should improve or decline at the same rate as (in this instance relatively narrowly based) standardised test outcomes concerned with reading skills or comprehension. We observed KS2 English test levels improving, proportionately, more than standardised reading scores. After controlling for improvements in Reading

Scale scores it appeared that KS2 English levels achieved by equivalent children rose every year except 1998 between 1996 and 2000, by about 0.1 of a level each year. This is in direct contrast to KS1, where national assessment reading and writing levels improved relatively slowly by comparison with standardised test data.

Earlier we produced scenarios where different views might be taken about the way improvements in learning might affect standardised test scores. We have explained how we can accept that some increase over time in the levels achieved via national tests, by 'equivalent' children, can be explained by the efforts of teachers and children bearing fruit. But we would have expected to have seen larger increases in the earlier years of the 1996 -2000 period than the latter ones, as teachers will have known what to expect in the KS2 tests and made substantial progress in adapting their teaching by about 1998 (when it appears that there was little improvement in test levels overall). But the rate of improvement does not seem to have slackened. Although this evidence is in itself necessarily inconclusive, the rises in KS2 English test levels compared to those in standardised test scores since 1998 were larger than those observed in KS2 Mathematics and might be consistent with the application of more lenient test standards.

There is also some evidence that there may have been gender effects within the shifts in test standards over time. Most notably it seems possible that the 1998 test might have been more lenient for girls than boys, with the reverse holding in 1999.

The only data concerning elements within the KS2 Tests available relates to Reading and Writing levels in 1998 and 1999 in one Authority. LEA 5 provided data regarding children's KS2 Reading and Writing levels from the KS2 English 3-5 Tests in 1998 and 1999. These supported two conclusions. Firstly, it was clear that any shifts in KS2 English test standards arose largely or wholly in the Reading component, as improvements in Writing levels were much lower and not inconsistent with some small genuine improvement in children's achievements. Secondly they suggested some gender variations; notably that boys appeared to gain more than girls in the Reading element in 1999. This does help explain how boys reversed the gender differences observed in 1998.

KS2 Mathematics

The (uncorroborated) data from LEA 6 suggested that children from the 1997 cohort (with equivalent aptitude test scores), on average, obtained better KS2 Mathematics results than those from the previous year. Modest improvements in achievement like this would not be surprising at this early stage in national testing and these data would not seem to be strong grounds for suggesting that test standards may have declined between 1996 and 1997.

However evidence from both LEA 1 and LEA 6 showed that children from their 1998 cohorts (again with equivalent standardised test scores to those from the previous year), on average, obtained worse results in the KS2 Mathematics test. This suggested that test standards became more severe in

1998. But we should note that 1998 saw the introduction of a new element in the KS2 Mathematics Test – Mental Arithmetic. In effect this introduced an additional hurdle for children to clear and it may well have been responsible for the fall in mean test levels. Does this change in the criteria mean that standards were made more severe? Certainly children were being asked to do more to reach a given level. But should their teachers not have been able to adjust their preparation to the new requirements? Unfortunately, when curricular/ test requirements are altered, teachers may need a short time to adjust. Results following major changes in examinations at 16+ in the last decade suggest that they appear to learn from their initial attempt to teach to a new scheme and children's performance is likely to improve in the second year of a revised form of assessment. The data from both LEA 1 and LEA 5 suggest that this might have been the case in KS2 Mathematics, as between 1998 and 2000 the net effect of changes in estimated mean KS2 test levels for children with equivalent standardised test scores in both these LEAs was a rise approximately equivalent to the fall observed in 1998.

This may well have left KS2 Mathematics test standards much as they were at the start of the period investigated.

Teacher Assessments
KS1 Aggregated Reading & Writing: Only one LEA's data provided evidence concerning the relationship between standardised reading scale scores and aggregated KS1 Teacher Assessments (TAs) for Reading and Writing. These were however large-scale annual datasets, relating to complete cohorts from each year. Teacher Assessments appear relatively stable. After controlling for substantial improvements in reading achievement it would seem that the average TA levels of children of equivalent (reading) ability had risen by only about 0.4 of a level between 1996 and 2000. Given that we might have expected the teachers to have made increasing (and increasingly effective) efforts to address the targets involved across this period, this would not seem unreasonable.

KS2 English: Teacher Assessments of KS2 English were available from three Authorities (LEAs 1, 5 and 6) and show quite similar year on year changes in TA levels relative to their children's standardised test data.

Year on year changes in KS2 English Teacher Assessments appear to have been less volatile than those in KS2 Test levels. It would seem that, after controlling for rising standardised test scores, TA levels rose a little in 1997, fell back marginally in 1998 and rose again between 1998 and 2000, finishing around 15% of a level above their 1996 starting point.

Improvement on this scale might be expected, given that teachers will have been refocussing their efforts to concentrate on the targets involved throughout this period.

KS2 Mathematics: KS2 Mathematics Teacher Assessments in LEA 1 were less volatile than test results in 1998, matching the upward trend in standardised test scores fairly well. In LEA 6, the boys in the 1998 cohort had

a slightly lower mean standardised test score than the 1997 cohort and their Teacher Assessments were also slightly down, whilst the opposite held for girls. But taken over the full period 1996 to 1998, changes in Mathematics TA results in this LEA were also a reasonable match to shifts in mean standardised test scores. Trends in Teacher Assessments for Mathematics at the end of KS2 would therefore seem better matched to the trends in standardised test scores than were the results from KS2 mathematics national tests. The evidence here suggests that KS2 Mathematics Teacher Assessments too are reasonably stable over time and yet sensitive to variations in achievement levels between cohorts.

The evidence regarding Teacher Assessments at both key stages is consistent; providing no reason to assert that teachers applied more lenient standards over the period 1996 to 2000. Indeed their assessments have been more stable over this period than the corresponding results from national tests - a finding few might have predicted beforehand. Teachers appear to have proved themselves capable of recognising the improvements in performance standards they have fostered in successive cohorts, without over-rewarding them. This was of course achieved in an assessment culture where teacher assessment co-existed with testing. The profession may in some part have taken test outcomes as a lead, assisting and/or governing their own perception of achievement standards, but Teacher Assessments have proved less volatile than test results.

Standardised test data's value for monitoring national test standards

There is thus no single trend in these data. Evidence from different key stages and curriculum areas points to different conclusions regarding national test standards. This evidence from LEA's standardised testing is anyway not in itself conclusive. It needs to be set alongside the data from the Project's experimental and other strands of work, to see whether or not the different sources and methods are broadly agreed.

But standardised testing data would seem to be of considerable interest in relation to national testing. It may well be worth pursuing this approach in future years, as an additional perspective on both performance standards and test standards which may assist those responsible for managing the national assessment system.

6.1 The Project's evidence: a summary

Experimental comparisons

Methodological innovation

The experimental design used an 'equivalent-groups' definition of comparability, whereby children allocated at random to equivalent groups taking two (present and past) versions of a test are expected to obtain similar results if standards are well aligned. It gets as close as is feasible to the obvious question - do today's children get the same results as if they had taken a test set x years ago?

We have discussed why it is in practice so difficult to address so obvious a question and, hence, why these were the first large-sample experimental comparisons of versions of large-scale UK public examinations or tests set in different years. Only small-scale studies and quasi-experimental comparisons, where children took a past version as well as their 'live' current test, have been reported previously.

We believe that we were as successful as possible in obtaining experimental conditions equally fair to both versions being compared. Conducting the research in Northern Ireland ensured that children involved were equally motivated to achieve on both versions. Furthermore, we tried to establish whether they might have been taught in ways which could bias the results. We looked at differences between the school and assessment systems and the curricula in England and NI; investigated the potential impact of these on performance on the questions set in the versions of the test being compared; and sought the children's teachers' views on the fairness of the comparisons. Few real threats to the validity of the comparisons were identified and any questions were recorded so that their potential effects on results could be considered.

The choice of 'medium-term' time intervals (of 3 to 5 years) between versions compared was probably a critical factor in this. As time goes by, curricular change makes it more likely that older versions of a test will become 'out of date' - i.e. less relevant to today's curriculum.

The design thus proved effective, with one reservation. The precaution we took of gathering some additional information about the calibre of the groups taking different forms of the test was advantageous, in that it provided an alternative analytic approach in a few instances where randomised matching by spiral allocation proved less than satisfactory, or, otherwise, provided a powerful analytic tool. Given the cost[1] of this kind of research, this precaution seems sensible if data from a suitable control variable can be found.

[1] For instance Project costs were in the region of £300,000, including contractual payments of £276,245 to UCLES, with the balance made up by printing and management costs incurred by QCA. The lion's share of these (say 80%) were incurred in conducting the experimental comparisons, across all subjects and key stages.

Key Stage 1

KS1 Level 2 Reading Comprehension test: 1996 v 1999
Taken at face value, the experimental comparisons suggested that the thresholds set for level 2c in the 1996 and 1999 versions of the Level 2 Reading Comprehension Test represented similar standards, but that there may have been a disparity between the two versions of the test in the sub-divisions of level 2 denoting higher achievement. Note however that the level 2b and 2a thresholds set in the 1999 version were stricter, rather than more lenient, than those in the 1996 test.

The 1999 test was in many ways an attractive one. It was liked by children and staff participating in this experiment, but it was based on a more complex text than the 1996 version and was undeniably more difficult. This is not to say that the developers of the 1999 test pitched it at too high a level of difficulty. Many children scored very high marks indeed on the 1996 version of the test, which was perhaps too easy to discriminate between the abler candidates as effectively as might be desired. But the concern here is with the cut-scores, not the difficulty level of the test materials. The 1999 cut-scores for levels 2a and 2b may have been a mark or two higher than was required to match the standards represented by their 1996 equivalents.

This validates the stability of the results reported nationally, at level 2c, between 1997 and 1999. It also suggests that the modest improvements reported at levels 2a and 2b (and thus involving a large proportion of the cohort) might be regarded as an under-estimate of KS1 children's improved performance on Reading Comprehension tests over this period. The evidence thus strongly supports a view that performance levels in schools have risen.

However, because the 1996 level 2 Reading Comprehension Test remained optional it did not play the same part in determining KS1 Reading levels as the 1999 version. Any conclusions are therefore limited to the comparison between the two versions of the test rather than the standards governing overall KS1 Reading assessments in 1996 and 1999.

KS1 Mathematics test: 1996 v 2000
The experimental evidence here reaches conclusions similar to those for KS1 Reading Comprehension. It suggests that the 2000 version of the KS1 Mathematics test seems to have applied standards for the award of level 2c which were at least as demanding as those applied in the 1996 version.

Moreover, higher up the ability range, for both the sub-divisions within level 2 and at level 3, the 2000 version's cut-scores appear to have made heavier demands on the children than the 1996 version's thresholds.

If the validity of this evidence is accepted, the improvements observed in results at a national level since 1996 seem more than merited - reflecting, indeed under-estimating, learning gains in schools. Confirmation that performance levels in schools at KS1 in Mathematics have risen over the period investigated, like those in Reading Comprehension at this key stage, is

in itself of considerable significance in light of the slow movements over time in standards in basic skills reported previously.

Key Stage 2

KS2 English: 1996 v 1999 and 1996 v 2000
The most important conclusion to be drawn from the experimental evidence regarding KS2 English, is that when considered alongside the fast improving distribution of KS2 English test levels over the period concerned, it is entirely consistent with the view that there has been a substantial improvement in children's performance. But these data do also suggest that national results might over-estimate the rate of progress.

1996 v 1999 The experimental comparison in the project's initial phase suggested that there may have been a shift towards more lenient standards in Key Stage 2 English test thresholds between the 1996 and 1999 versions, especially at levels 4 and 5. Children taking the 1999 version obtained average levels significantly better than those taking the 1996 test and adjustments in the level 4 and 5 cut-scores of up to five marks appeared to be required to equate the two versions.

Because the criteria for awarding Writing marks (and the marks awarded to the groups taking the two versions) were essentially unchanged, we could see how standards in the 1996 and 1999 versions of the test diverged, despite the lack of separate cut-scores for Reading and Writing in 1996. Average total marks on the two versions differed by about four marks, with higher marks for the 1996 version. Aggregated Writing marks were the same on both versions and the four mark difference in average marks was thus attributable to the 'harder' 1999 Reading test. But whilst Reading (and total) marks on the 1999 version were, on average only four marks lower, the (overall) 1999 cut-scores for levels 4 and 5 were nine marks below those applied in 1996; over-compensating for the difference in (Reading) test difficulty.

1996 v 2000: The Project's replication of the initial experimental comparisons in KS2 English contrasted the 1996 version with 2000's test instead of 1999's and provided an opportunity to find out if the experimental methodology was robust. Could a replication produce consistent results?

The answer was, most emphatically, yes. The second study replicated the earlier findings in some detail, supporting their joint conclusions. If anything the gap between the standards applied in the 2000 and 1996 versions was marginally wider than that observed a year earlier; notably at the level 5 threshold - where national results in 2000 had in fact improved somewhat. Thus the 2000 cut-scores also appeared relatively lenient, especially at levels 4 and 5 - where increases of 5 and 7 marks respectively were required to equate to 1996.

As previously the disparity seemed to arise largely from the Reading component in the tests. Writing marks obtained on the two versions were much the same. This matches the national pattern of results over time.

Within KS2 Writing, the standards applied through the various writing task options (together with their marking schemes and cut-scores) available within the 1996 and 2000 versions seemed fairly consistent. Just as was observed in the earlier comparison with 1999, setters and markers balanced the optional tasks reasonably fairly, so that children of similar ability choosing different tasks obtained approximately equivalent marks.

But we cannot be sure that the assessment of writing at KS2 is entirely unproblematic. There were hints in the data that children attempting writing options with more clearly defined and/or structured tasks might have obtained slightly better marks. The provision of such support within the question paper has become more prevalent over recent years and may be making writing more accessible. We should be cautious here as other experimental research in progress indicates that such scaffolding may not always provide the benefits expected (Green, 2001) but further work may be desirable. There is also the potential for variation in the 'difficulty level' of the spelling test, which is subsumed within Writing at KS2, as variations in the difficulty of spellings could mask shifts in either attainment or marking standards in writing.

The development team responsible for the KS2 English tests (like those producing English tests for KS1 and KS3) faced particular challenges. For instance, where sets of questions depend on a common stimulus (such as a passage in a reading test), the stimuli's complexity and features govern the difficulty of the whole set of questions. This limits the options available in test construction and makes it more difficult to avoid fluctuations in the difficulty level of the tests. To compensate, cut-scores must then be relatively mobile from year to year too; making the annual equating and standard setting process especially demanding.

The improvement in national results over this period is thought to stem largely from improved performance in Reading. If valid, our experimental data question the extent[2] of this improvement.

KS2 Mathematics: 1996 v 1999

In the experimental comparison of Key Stage 2 Mathematics there was no indication of any disparity in standards between the 1996 and 1999 versions of the test, despite the potential for disturbance by the introduction of the mental arithmetic element between these dates. Small observed differences in levels achieved via the two versions proved insignificant once we had controlled for variation in the ability of boys assigned to the two experimental groups.

[2] Increasing the level 4 threshold (set near the 70th percentile) by 5 marks (about one third of a standard deviation) in 2000 would have reduced the number of children reaching this level overall. The assumption that the distribution of KS2 English test marks is approximately normal would lead to an estimated effect size of the order of 10%.

The experimental evidence therefore provides no reason to challenge the validity of the improvements in KS2 Mathematics national test results reported between 1996 and 1999.

KS2 Science: 1996 v 2001
Again, perhaps the most important feature of these experimental data is that it provides support for the view that there has been a great improvement in children's performance on KS2 Science tests in recent years. But as in KS2 English, there are signs that a small part of the very large improvement in national test results reported between 1996 and 2001 may be a product of a shift in test standards.

Our experimental comparisons showed better results being obtained by children who took the 2001 versions of the KS2 Science test. The size of the disparity varied across the range of ability. Weaker children derived a greater advantage from being assigned to the 2001 version (perhaps a third of a level but the ablest were hardly affected at all. Curricular factors suggested that children in NI might have been at some relative disadvantage on the 1996 version as compared with 2001's, making these conclusions conservative.

Equating suggested that to align the 2001 test with the standards set in the 1996 version, the level 2 and 3 thresholds should have been two marks higher and the level 4 threshold four marks higher; although the level 5 threshold would have remained unchanged. Without access to operational mark distributions we can only estimate the numbers of children likely to have been affected by such changes. But adjustments of this order[3] would probably have cancelled out only a small proportion of the huge improvements in results on KS2 Science tests observed between 1996 and 2001. There is no evidence in these data which bears upon when or how any such shifts in Science test standards might have occurred, although it might not be unreasonable to point to the substantial improvements in results in 1997 and/or 2000.

Key Stage 3

KS3 English: 1996 v 2001
KS3 English required a methodological adaptation to cope with the impracticability of including Paper 2 (based on different selections of Shakespeare plays) in an experimental comparison of the 1996 and 2001 tests. Instead, children completed the 1996 and 2001 versions of Paper 1 only, and the statistical relationships between (boys and girls) Paper 1 and Paper 2 scores in final trials data for both versions (supplied by the KS3 English Test Development Agency) were used to generate predicted Paper 2 scores and test levels. Comparing these revealed that slightly higher predicted levels were obtained by the group assigned to the 2001 test.

However a small-scale investigation of possible drift in marking practices was mounted by re-marking a small sample of 'archived' 1996 Paper 1 scripts.

[3] Perhaps reducing the percentage of children gaining level 4 in 2001 by around 5%.

This suggested that, as hypothesised, curriculum and learning advances since 1996 may well have resulted in today's markers' expectations being higher than those of 1996. Given that contemporary markers were responsible for marking both the 1996 and 2001 versions in our experimental comparison, this appeared to explain the differences in predicted levels observed. It seemed best to conclude that overall KS3 English test standards have been successfully maintained over the period 1996 to 2001. There is no basis here on which to challenge the improvements in KS3 English levels recorded nationally; again an important conclusion.

Investigation of the equivalence of the marking standards applied to optional writing tasks in both the 1996 and 2001 versions revealed that parity was achieved and any unfairness to the groups of children selecting the various options available was avoided.

KS3 Mathematics: 1996 v 2000

In Mathematics at KS3 the experimental evidence showed that pupils taking the later (2000) version of the test obtained better results (in terms of levels awarded) than those allocated to the 1996 version. The tiering arrangements within these KS3 Mathematics tests complicated comparisons. In essence, relative lenience in the 2000 version is greater for lower ability pupils. Those taking Tier 3-5 of 2000 achieved about 50% of a level better than equivalent pupils taking the 1996 version; those taking Tiers 4-6 and 5-7 about 25% of a level better; whilst the ablest, taking Tier 6-8, achieved only about 10% of a level better. Assuming the comparisons made to be valid, many of the 2000 version's cut-scores appeared to require substantial upward revision to bring them into line with the 1996 version. Equated cut-scores for each level within each tier[4] were detailed in section 2.8.

The format of the KS3 Mathematics test was 'disturbed' by the introduction of a test of mental arithmetic from 1998 onwards. These data suggest that some of the disparity in standards between the two versions of the tests relates to this new element, most notably in the lower range of attainment. But far from all the variation in average scores between versions was accounted for by the inclusion of mental arithmetic: only about a 0.2 of a level in T3-5 - less than half the overall disparity between versions; 0.17 of a level (more for boys, less for girls) in T4-6 - representing about two-thirds of the overall disparity between the 1996 and 2000 versions; and 0.06 of a level (again more for boys than girls) in T5-7 - just over half the overall difference. In T6-8 pupils did not gain from the inclusion of mental arithmetic. The initial 'component thresholds' for mental arithmetic may well have been too low, especially at lower levels.

KS3 Science: 1996 v 2001

Experimental comparison of the levels achieved, over all tiers, via the 1996 and 2001 versions of KS3 Science, suggested that after controlling for variations in ability arising from gender effects and assignment to the groups

[4] Tiers make it especially difficult to quantify the effects of changes of thresholds. However in this case access to distributional data from a national sample provided by QCA enabled us to estimate that the overall effect of applying the 'equated' thresholds in 2000 would probably have been to reduce the percentage of children reaching levels 5 and 6 by around 8% in each case.

taking the two versions, the levels achieved were very similar. These data therefore suggest that the quite substantial gains in KS3 Science test results reported nationally between 1996 and 2001 were merited; reflecting improvements in teaching and learning in schools. Like the similar conclusions regarding almost all the curriculum areas at all three key stages investigated, perhaps this should be recognised as the most important inference we have been able to make.

Further analyses explored the equivalence of levels achieved on the 1996 and 2001 versions of the test via the two tiers, targeted at levels 3-6 and 5-7 respectively. In essence these confirmed the overall verdict that test standards had remained well aligned over the period 1996 to 2000, although it appeared possible that the 2001 thresholds for Tier 3-6 might have been a mark or so too high.

Evidence from LEA's standardised testing programmes

Although Local Education Authorities proved willing to help wherever they could, it was surprisingly difficult to find LEAs with databases where standardised testing data could be linked to information concerning children's achievements in national assessments. Only six authorities were identified who could provide such data. These used a variety of different standardised tests administered at different stages in the school year, often for different sequences of years. Analysis of these data therefore involved a series of case studies, essentially comparing the rate of change in average KS1/KS2 levels achieved to the rate of change in standardised test scores. Behind these comparisons is an 'equal-calibre' definition of equivalence. Comparability is achieved if children of equivalent calibre (measured by their standardised test scores) obtain similar levels in different years. Overlaps between the sequences of years involved provided some opportunities to confirm evidence from one Authority's data by replication in similar data from others.

Very significantly, there were widespread indications from the various LEAs that standardised test scores have risen between 1996 and 2000. This was true for standardised tests of reading ability and reading comprehension at KS1 and KS2, and, quite possibly, for broadly based standardised mathematics tests at KS2. There seems no reason why teachers should have increased their efforts to prepare children for such tests – and some reasons why they might have taken less interest in doing so over these years. The evidence therefore strongly suggests that schools have produced genuine gains in achievement in recent years. Whilst gains were not universal, those observed were sometimes very substantial.

- At KS1, there was sound evidence for an improvement in reading scale scores. An improvement of about 25 % of a standard deviation over the four years 1996 - 2000 seems a reasonable estimate.

- Whilst the improvement in reading test scores at KS2 was probably less dramatic, substantial improvements were seen in the two LEAs who could

provide sound evidence of improvements (up to 16% of a standard deviation over four years). Overall, an estimated improvement of 10% of a standard deviation between 1996 and 2000 would seem a fair compromise between the changes observed in the four Authorities providing such data.

- Two LEAs were able to provide data for large scale use of standardised tests of mathematics at KS2. Unfortunately neither set of data included the whole of each cohort and changes in the numbers of children involved in testing made the data less than ideal for estimating changes in performance levels over time. However the evidence available did suggest that standardised test scores in mathematics at KS2 may also have risen substantially between 1996 and 2000 - perhaps by around 25% of a standard deviation.

Clearly, any such improvement in achievement validates some gains in national assessment results. These have indeed improved too, in all the LEAs providing standardised test data, but the matter cannot simply be left there. What can these data tell us about the size of improvements in national assessment results we might regard as justifiable? Using standardised tests as a common yardstick to compare national assessment standards in different years involves an implicit assumption that if children's achievement is changing over time, standardised test scores and national assessment level distributions should change at a similar rate, so that, on average, children with equivalent test scores in different years achieve the same levels. This is not logically sustainable, as there are reasons why we might expect teachers to have switched efforts away from the things tested in these standardised tests and towards the domains addressed by national assessment. Whilst this line of argument might be advanced to explain national test results improving more quickly than standardised test scores, it is difficult to envisage valid reasons why the opposite might be the case, which would suggest that national tests etc. might have become more severe.

- This opposite case is just what is found in evidence relating standardised reading test scores to overall KS1 Reading levels. Two LEAs were able to supply sound data for this and, once we had controlled for improving Reading Scale scores, it seemed that equivalent children were receiving worse results in 1998 than in 1996 and 1997. Moreover, despite a recovery in 1999, evidence from 2000 suggested that levels received then were back to 1998's low point. It appears safe to conclude that the standards which applied in overall KS1 Reading assessments in 1996 have at least been maintained and may have become a little more severe by 2000. Sound evidence from one LEA relating to the KS1 Reading Task suggested that the same conclusion holds for this 'component' in KS1 Reading, where levels awarded to equivalent children fell in both 1997 and 1998 and only partially recovered thereafter. Might teachers have become more conservative in awarding KS1 Reading Task levels since the Reading Comprehension Test became compulsory? Data regarding performance on the KS1 Level 2 Reading Comprehension test suggested that test levels for children of equivalent reading achievement were on average a little higher in 1998 and 2000 than they had been in 1997, the

year in which this test was first made compulsory. But the size of the changes observed were small enough to be consistent with the kind of improvements which might have been expected as teachers got to grips with preparing children for this new test - and hence with test standards having been maintained. This is in keeping with the evidence from the Project's experimental comparisons involving this test.

- Evidence from the same two LEAs was remarkably consistent in suggesting that average KS1 Writing levels for children of equivalent reading ability may have fallen in 1997, and again in 1998, although some variation by gender appears possible. A partial recovery in 1999/2000 perhaps failed to recover all the ground lost earlier, which leads to the conclusion that KS1 Writing assessments may have become more severe since 1996. The criteria for the award of levels has not changed, but might the introduction of sub-divisions of level 2 in 1996 and the provision of new exemplar materials since then have discouraged some teachers from awarding level 3? Analysis of data relating to the KS1 Spelling test shows that children of equivalent reading achievement gained slightly higher levels, on average, in 2000 than in 1997. Again however the 'improvements' were quite small and were consistent with teachers having directed more attention to teaching and learning in this area. The data available therefore do not provide grounds for suggesting that KS1 Spelling test standards had become relatively lenient.

- In contrast to the pattern observed at KS1, KS2 English test levels improved more quickly than the children's standardised test scores. The evidence would seem to suggest that after having controlled for the effects of improving standardised test scores, 'equivalent' children obtained levels which rose, on average, by about 0.1 of a level each year (except 1998) between 1996 and 2000. This rate of improvement is higher and more sustained than that observed in KS2 Mathematics (see below). Whilst some improvement might have been expected, especially in the early period when teachers were familiarising themselves with KS2 English tests, these data might lead to concern that the KS2 English tests may have become more lenient over time. As such, this standardised testing data fully supports the findings of the Project's experimental comparisons in KS2 English, which suggested that about half the improvements in test results between 1996 and 2000 may have resulted from changes in test standards. The standardised test data also provided indications of gender differences in the course of shifts in standards. Boys may have lost ground to girls in 1998 but gained it back in 1999, largely because of their performance on the reading element within the test. These standardised test data confirm that the Reading element in the KS2 English test was the source of any disparities in test standards over time, as the experimental comparisons had also suggested. Improvements in KS2 Writing levels (for children of equivalent ability) were not too large to be plausible, given that improvements in schools' and children's performance might have been expected as teachers' re-directed their efforts towards new targets.

- Overall, there was no evidence of a shift in KS2 Mathematics test standards in the period 1996 to 2000. However counter-balanced shifts occurred within the period, which indicates that even Mathematics is vulnerable to instability in standards. The evidence suggested that the 1998 KS2 Mathematics Test was more severe than the 1997 version. The introduction of mental arithmetic in 1998 probably explains why; by providing a new set of demands on children. By 2000 it would appear that the pendulum had swung back, perhaps because teachers had adjusted teaching and learning programmes to meet these new demands: so that over the whole of the period 1997 to 2000, changes in KS2 Mathematics Test levels were in keeping with those in standardised test scores. Yet again this evidence matches that from our experimental comparisons, which produced no suggestion of changing standards in KS2 Mathematics. But it is worth noting that with the advent of mental arithmetic, the children are being asked to do more now.

- Interestingly, several LEAs supplied data for teacher assessments as well as test levels. These related to Aggregated Reading and Writing at key stage 1 and to both English and Mathematics at key stage 2. In all cases, these appeared to be rather less volatile than test results and there was nothing to suggest that teachers were becoming more generous over time than might have been warranted.

Teachers' judgements about 1996 and 1999 KS2 English scripts

Here another concept of comparability was employed, whereby children whose work (in response to different versions of the test) was judged (on balance) to be of equal worth were expected to gain similar national curriculum levels. This 'equal worth' definition embodies the teachers' judgemental trade-off, balancing relative task complexity against the frequency of correct / high quality answers. We should note too that judgemental comparisons of standards across different tasks are necessarily imprecise, making the detection of small differences unlikely.

Whilst teachers given access to KS2 English scripts representing key mark points in the 1996 and 1999 versions of the tests were not unanimous, the majority view supported the empirical evidence. The work of children on the 1996 level 4 cut-score mark was judged, on balance, superior to that of children at the minimum mark for the award of Level 4 in 1999. The teachers' judgements thus supported the conclusions suggested by the experimental comparisons and the LEA standardised testing programme evidence in this subject at KS2.

Children's perceptions of evolving features in national tests

It is important that children's perceptions are taken into account, since they are the ones taking the tests. Yet their opinions are rarely sought. The children interviewed for this study were perceptive and were able to

distinguish a range of features in the test materials and to express their opinions effectively, even those as young as seven. The modified version of Kelly's repertory grid questioning technique enabled the children to verbalise their own thoughts and identify salient features of the materials, allowing elicitation of their personal constructs and perceptions.

A range of salient features and reasons for their preferences between them were identified by children, with some common between subjects and key stages and others subject and/or key stage specific. Children were able to discuss the materials and raised valid and interesting points.

Overall, the materials in each 'paired comparison' taken from more recent tests (1999, 2000 or 2001 as opposed to 1996) tended to be preferred. There were exceptions, such as the coloured advertisement used in 1996 KS3 English. Qualitative data like these cannot be regarded as conclusive, as the materials used here only sampled the test materials concerned and the views of quite small groups of children were canvassed. But the children's opinions are convincing and strongly suggest that test developers' efforts to make test materials more attractive and user-friendly have been successful.

We cannot say how far such changes might affect performance. Those which make it easier for children to understand what they are asked to do, or simplify the ways in which they are asked to respond in order to demonstrate what they know, would seem likely to enable more children to demonstrate competence. Other features may engage or motivate children and in doing so may lead them to better performance, but engagement does not necessarily mean that children will produce the responses markers seek. We can only say that the children seemed to think that many of the features which have been introduced in more recent versions of national tests were interesting and/or attractive and often considered them likely to help them negotiate the tasks set.

Children's perceptions are important, not only because of the potential to improve performance, but also because potential negative perceptions and side effects should be avoided if possible. One such side effect is the stress felt by children undergoing high stakes assessments. Insight into children's reactions to test materials could inform current thinking on test anxiety. Reay and Wiliam (1999) describe how 'children are simultaneously active in the assessment process and profoundly affected by it... there are strong currents of fear and anxiety permeating children's relationships to the SATs process'. Across the key stages, children interviewed in this Project responded positively about being asked to take national tests. Younger children had some difficulty with the question but tended to feel that the tests encouraged them to learn. For older children, reasons given often related to establishing levels for their next school or for examination setting. Some KS3 pupils disagreed about the importance of the tests, feeling that they were for the school's benefit rather than the pupil's. Older children also commented on negative side effects, such as time lost to 'revision'. Most felt anxious before live testing and relieved afterwards, although the general feeling was that the tests had not been as difficult as anticipated. Some anxiety was due to 'scare-

mongering' by other pupils. Direct feedback from children who have taken the tests previously, and felt them less frightening than expected, could help allay the pre-test fears of others. For older children it could be prudent to clarify the purpose of the tests, to reduce confusion and stress.

The efforts to make tests clearer to understand and easier to respond to, as considered in this strand of research, were welcomed by children and must have helped to soothe their initial anxieties. Further efforts to improve them remain desirable.

But this raises interesting questions regarding test standards. The improvements in test materials were perceived by children to make the tests more accessible. Do such features really make tests easier[5]? If so should test thresholds be adjusted, over time, to compensate for this, so that today's children obtain results no better than they might have received had they sat yesterday's tests? Or should such developments, helping children to show what they can do, be regarded as a valid means of recognising (and perhaps even improving) performance which should be both welcomed and reflected in better results?

Questions like these take us back to the definitions of comparability underpinning our empirical investigations. Some of the more inviting answers would indeed attack the very foundations of the definitions used; undermining their null hypothesis - that we should expect children of an equivalent level of achievement to obtain similar results on national tests from different eras.

If we believe we can demonstrate that national tests have been 'improved' in ways which improve accessibility, as we do; does this mean that we might expect to find that equivalent children taking more recent versions of national tests obtain better results? This was of course just what we found in English and Science at KS2 and in Mathematics at KS3, but not in the other subjects at these key stages. It would be hard to separate the effects of increased accessibility from other factors, especially schools' efforts to improve learning. Note too that any such expectations must make us even more surprised if national tests appear to have become more severe over time.

Is there any overall pattern?

Where variations in test standards across years seem apparent, accessibility issues like those discussed above might question which version (earlier or later) of a test set the most appropriate standard. But let us for now make the assumption (one integral to the use of test scores to support policy making by monitoring performance) that we expect pupils with similar achievement to obtain equivalent results in different years. Can we draw together what the project's various sources of data can tell us?

[5] Research suggests that the factors affecting difficulty are complex and the 'later' tests perceived by children to be friendlier were sometimes more difficult, in terms of average marks.

Children's performance has improved
The feature of these data which must be highlighted before all else concerns the support they give to the inference that performance levels in all subjects at all three key stages have risen over the period concerned. Taken in conjunction with the changes observed in national test levels since 1996, the experimental data suggested that the national curriculum and its associated testing regime had been successful - through the efforts of schools, teachers and children - in levering up standards of attainment in England. The data from LEA's standardised testing programmes provided powerful support in this respect; cross-validating the experimental evidence wherever both sources of information were available to the project.

Test standards have been maintained in most subjects/key stages
At KS1 experimental comparisons were consistent across the two subjects involved. Comparisons of both the Level 2 Reading Comprehension (L2 RC) test and the Mathematics test suggested that the 'later' versions (1999 in the case of L2 RC and 2000 in the case of Mathematics) were well aligned with the standards set in the 1996 versions at the 'key' level 2c threshold. But at the higher (2b and 2a) thresholds the standards set in later tests in both these curriculum areas were more severe than those applied in 1996. The improvement in KS1 test results nationally over the periods in question has been quite modest. Between 1997 (when L2 RC tests became statutory) and 1999 the percentage of the cohort achieving level 2 in the Reading Comprehension tests rose by only 5%, from 74 to 79%, whilst the proportion reaching level 2a or higher remained stable. Our experimental findings suggest the modest improvement at 2c was valid and that a similar improvement higher in the ability range might also have been merited. The evidence relating LEA standardised testing data to KS1 RC test levels was entirely consistent with this conclusion. Nationally, KS1 Mathematics test results also showed fairly modest improvements between 1996 and 2000: the percentage of each cohort reaching level 2 improved by 8% (from a high baseline of 82%), whilst the percentage reaching level 2a rose by 7%. Again the experimental data validates the modest improvement in results at level 2 and suggests that even more might have been merited at higher sub-levels.

But at KS2 the experimental outcomes were not so consistent across subjects. In Mathematics the experimental evidence suggested that test standards had remained constant between 1996 and 1999, apparently validating the substantial improvement nationally in the percentage reaching level 4 of 15% (from 54% to 69%) between these years. Again the evidence relating LEA standardised testing data to national test results was consistent with this conclusion; suggesting that despite some disturbance to standards when mental arithmetic was first introduced in 1998, re-adjustments since then have kept standards in line with the past.

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. It should be noted that even if it is accepted that our evidence indicates a change in standards between 1996 and 1999 / 2000, this does not conflict with the findings of the Rose Panel.

The Panel held that the procedures for setting standards in 1999 had been followed correctly and were by and large adequate and had no reason to doubt that the standards set were in line with those from 1998. Our data do not bear upon procedural matters, let alone dispute the Panel's conclusions in this respect. Nor do our data say that 1999's KS2 English standards were lower than those of 1998, the year on year issue with which the panel was concerned. Our experimental evidence that standards in the 1999 version of KS2 English differed from those in the 1996 version does not tell us when the change happened. Movement may equally well have occurred between 1996 and 1997, or between 1997 and 1998. Or a series of incremental changes between successive versions of the test – each in itself difficult to detect - might have taken place. The LEA standardised testing data available led to conclusions largely consistent with this latter suggestion, as KS2 English test levels for children with equivalent standardised test scores rose by about 0.1 of a level each year between 1996 and 2000 - except for 1998, when overall standards appear to have remained at their 1997 level, although boys lost ground relative to girls then (and regained some of it in 1999). The standardised test data also confirm the evidence from the experimental comparisons, 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.

In KS2 Science too the experimental evidence indicated some differences in test standards; but insufficient to invalidate most of the improvements in national test levels recorded between 1996 and 2001 - during which the percentage reaching level 4 rose by 25% nationally.

At KS3 experimental outcomes suggested that test standards had remained constant over the period 1996 to 2001 in both English and Science. Both these had seen only modest improvements in national results by comparison with KS2's large gains: between 1996 and 2001 the percentage reaching level 5 in English nationally increased by only 7% (57% - 64%), whilst that in Science rose 10%, from 56% to 66%. The improvements in national results for Mathematics were of a similar order, with the percentage reaching level 5 also rising 10% from 56% to 66%, but the experimental evidence in KS3 Mathematics may bring this into question. However the 'mid-term' introduction of mental arithmetic to the testing regime undoubtedly created a discontinuity which contributed to the problems involved in setting consistent KS3 Mathematics test standards.

So, at KS1 both tests investigated seem to have become more severe; at KS2 greater leniency seemed likely in two (English and Science) out of the three curriculum areas, whilst standards were maintained in Mathematics; but at KS3 standards appeared more lenient in Mathematics, whilst being maintained in the other two subjects - English and Science. This is hardly a uniform pattern, which in itself contradicts any suggestion that there may have been a concerted effort to manipulate national test standards in such a way as to improve results.

6.2 Discussion

Diversity and cohesion

The summary of the Project's evidence above describes how our use of a variety of methods and data for investigating comparability over time produced diverse results. In some cases standards over time seemed in line, in others tests had become more lenient or more severe. But within each of the key stages and subjects investigated the different sets of evidence were remarkably cohesive. Where relevant data from LEA's standardised testing programmes were available (i.e. Key Stage 1 and 2 Reading/English and Mathematics) to supplement our experimental comparisons, they pointed to similar conclusions, as did the qualitative judgements of teachers looking at borderline scripts from 1996 and 1999 KS2 English.

The Project's strategy of cross-validation through different methodologies thus proved successful. Direct experimental comparisons of the test forms from different years were effective and investigations of their validity in a Northern Ireland context were re-assuring - although comparisons outside the target population would be better avoided if possible. The use of LEA standardised testing data helped to provide a broader picture, which seemed to confirm the experimental comparisons and helped fill some of the gaps. Both strands of qualitative work built on these empirical comparisons. Teacher's judgements helped to confirm the experimental evidence in KS2 English and the children's perspective extended our understanding of the tests and may have helped to explain the ways in which they are evolving. These all employ different perspectives, helping us to appreciate the variety of concepts of standards underpinning comparisons of standards within the operation of our testing system. In a field like this, whether for operational or research purposes, several perspectives on the problem will often be safer than one.

Can we accept the empirical evidence at face value?

Given that in each case the available sources of evidence all tend to point in the same direction, can we accept the apparent variations in standards (as summarised in 7.1 above for each key stage / curriculum area) at face value? Or might there be other valid explanations where standards appear to have shifted?

The definitions of equivalence behind the evidence
Remember the variety of definitions of equivalence and associated assumptions that were in use. Experimental comparisons expected equivalent groups of children to obtain similar results from different versions of a test, whilst the use of standardised test data assumed that children of equal calibre, as measured by some common 'monitor' variable, should obtain the same result if they took different versions of the tests. These are not widely dissimilar and may seem self-evident and fair, but very different definitions may be supportable too. For instance the teachers' comparisons of scripts from different versions required them to judge if sets of scripts representing

points on the mark scales from different years were of equal worth; having traded-off the quality of the answers against the difficulty of the different tasks in the tests being compared. This is not unlike the ways in which teachers and examiners are asked to contribute to the process of setting standards each year, behaving like connoisseurs (Sadler, 1987) who are able to sense the relative value of diverse achievements. For equivalence here, children judged to have produced work of equivalent merit (however imprecise the criteria and procedure for this) should receive the same rewards from different versions of a test. This approach has been described (e.g. Cresswell, 1994) as criterion-referenced, although such holistic judgements are quite different from attempts to specify, apply and aggregate sets of judgemental criteria, which Cresswell (2000) described as 'strongly criterion referenced' - whilst pointing out that this approach has proved too inflexible to be of use in school examinations.

Accessibility, criterion-recognition and equivalence
But we have already suggested that aspects of this need to 'trade-off' the observed quality of work against task variety and complexity to make comparisons between different tests raises interesting philosophic issues. Some features of national tests have been changed quite deliberately between 1996 and 1999 as test developers have rightly been urged to learn from schools' criticisms of earlier tests. Such improvements aim to make testing a more interesting experience, to motivate children to do their very best and to make the tests as fair as possible by making the tasks clearer so that children can demonstrate their achievements. Our qualitative investigation of children's perceptions of national tests (at all key stages and in all curriculum areas) shows how children themselves can recognise the features introduced to help them and value the efforts which have been made to make tests more interesting and appealing. Our national tests are now amongst the best in the world in terms of their clarity and design and user-friendliness. As such progress in test design is realised, so enhancing accessibility, should we then not expect more children to be likely to 'do well'? Greater accessibility 'should' mean that children are more likely to be able to show what they know and can do, thus enhancing test validity by avoiding 'false-negative' results - i.e. failing to recognise achievement children actually possess.

There is an inherent tension here between the desire to improve the quality of tests as our (as yet very young) national assessment system evolves and the need to incorporate year on year equivalence, so that we can chart progress in learning in schools. If we change the tests, so that more children are able to demonstrate that they have the qualities we seek to reward, does this not also change test standards?

Could we live with a testing system where equivalent children would get different results in different years as accessibility varied? Many might view this as only right and just, arguing that if the required knowledge etc can be recognised in children's responses to test questions it ought to be rewarded. After all, if we failed to do so as effectively in earlier years is this not our failure (not the children's) which should be corrected? We might describe this

as a ‘criterion-recognition’ view of standards, where evidence that a desired criterion has been achieved yields the same result, even if changes between versions of the tests has affected the likelihood that children can produce it.

Certainly no one has opposed enhancements to the accessibility of national tests and many of those responsible for teaching and learning may have some sympathy for this view. Indeed qualitative value judgements which in various ways contribute to standard setting in national tests are sufficiently inexplicit that this point of view may have featured in some of the decisions on test thresholds in recent years. If so, it may have contributed to some of the apparent variations in standards over time our empirical comparisons have detected.

But the criterion-recognition perspective on standards shares some of the more seductive characteristics of an overly simplistic approach to criterion-referenced testing and contains the same key flaw. Adopting the criterion-recognition model depends upon judgemental approaches to standard-setting. Even if these could be made consistent (and this is unlikely) it would produce inherently unstable results. Improvements in accessibility would result in more children displaying the required knowledge and, hence, improved results, year on year. In contrast, poorer results would be called for if (presumably by mistake) a new version of a test included relatively complex questions or stimuli and consequently children’s work was less likely to contain the characteristics wanted. If test development changes are designed to enhance ‘accessibility’, we might expect this to lead to a drift towards ‘better’ results year on year. But the empirical evidence regarding KS1 tests in this study amply demonstrates that this will not always be so. We recognised at the outset that it is simply not feasible to develop a series of tests without unpredictable variation in these respects. However a national testing system where results fluctuate uncontrollably, without reference to changes in the quality of teaching and learning, would be unsatisfactory, because of the national monitoring function.

But though we are confident that the criterion-recognition perspective would not provide an effective basis for maintaining national test standards, it cannot simply be dismissed. It has informed our thinking and may help to explain how we have arrived at current test standards. The validity of its contribution is a legitimate matter for discussion as we decide how standards are to be determined in future, as must be the need to control its effects.

The need to control change

Changes in the curriculum assessed by tests and in the tests' own features make it much more difficult to set equivalent standards year on year. Our infant national testing system has seen frequent changes in both these regards but we would argue that the national curriculum in England is now reaching a more mature stage, where stability is more important to the quality of teaching and learning than further refinements to what should be taught. It must also be recognised, quite explicitly, that there is a need to provide a

stable basis for assessments over medium-term intervals if their system monitoring function is to be effective. Otherwise national tests will not provide a sound basis for policy making. In the long-term, rigidity in the content and style of the tests is undesirable, but control over curriculum renewal and the continued enhancement of test accessibility, involving some loss of flexibility in the short-term, may help ensure stable test standards.

In conclusion

On balance, despite these reservations regarding the potential and legitimacy of the effects of enhanced accessibility on test standards in recent years, we believe that empirical evidence we have gathered concerning medium-term changes in test standards is a valid contribution to the debate. The evidence shows how difficult it is to determine standards and gives the lie to any theory of conspiracy to undermine them. It would seem that between 1996 and 2000 standards in the components of the national testing system were maintained, or indeed became marginally stricter, more often than they can be challenged.

6.3 Policy recommendations

A cyclical approach to curriculum & assessment system renewal

Having succeeded in helping to lever up standards of performance in schools, England's national testing system is perhaps reaching a state of maturity in which we might usefully manage progress and change more explicitly, in the interests of better measurement. Management of the review and renewal of the national curriculum has already recognised that teachers would appreciate greater stability and moved towards a five-year cycle (Colwill, 1997). We recommend that this should receive new emphasis and that improvements to the allied testing system should be integrated within this cycle.

Cresswell (2000) describes very clearly how shifts in the curricular basis on which children are to be compared make quantitative comparisons of performance (or equating) over time impossible. These theoretical problems must be acknowledged before we can begin, pragmatically, to solve the measurement problems involved. Creating medium-term stability in the curricular and assessment regimes is an *essential* prerequisite to the maintenance of test standards.

A pro-active approach to the management of changes to the curriculum and national tests will be necessary, with any desired changes to both being 'developed', ready to be implemented at intervals of several years (about five seems reasonable) instead of introducing them piecemeal. Schools and teachers would doubtless welcome such relative stability, although politics' penchant for immediate action would be frustrated.

Only if this stability can be attained can we hope to devise reasonably effective objective procedures for equating tests within each cycle. By also managing the transitions from one cycle to the next conservatively, we should then be able to collect national test data fit to inform policy making.

A baseline equating strategy

The current focus on year on year equivalence is an inherently weak strategy, in which the dangers of incremental drift in standards are readily apparent. Given medium-term curricular/assessment stability, we would recommend switching the focus of test equating, away from equivalence year to year, to a stepwise approach involving equivalence between a series of successive years and a 'stable' baseline before moving (at the transition between curricular cycles) cautiously to a new baseline, assuming that curricular changes then require it. This is the key to significant improvements in the quality of test equating possible, by comparison with current arrangements.

What form should baselines take?

A recent innovation in the national testing system should provide the ideal baselines for each curriculum area / key stage. Test development schedules have wisely incorporated the production of 'reserve tests', incorporating

revisions to assessment arrangements due to take effect from 2003, alongside the development of operational tests for each successive year. These are an insurance, providing an alternative test ready for use should there be a breach of security. Like the operational tests the reserves will be kept strictly secure and it is to be hoped that they will never be called upon. The reserves will be developed alongside the 2003 versions of each national test. As full-length parallel[1] tests, future versions could be equated to these very effectively during their final trials, conducted (as is customary) under highly secure conditions, as described below. The reserve tests could from then on double as baseline instruments.

Final equating trials

'Final equating trials' could be built into test development schedules at little or no extra cost[2], about twelve months before each test's operational use, after the final version of each test had been fully and finally cleared by all government agencies[3]. This would overcome two major weaknesses in current procedure: the imponderable effects of post equating changes required to the test and the unequal motivation of participants in current pre-tests towards the future test concerned and their own live version, to which it is equated. In the proposed trials, children would be taking either the baseline instrument or a future version of a test. They would not know which and would anyway have no reason to be more highly motivated on one than the other.

These equating trials should employ the same experimental design as the present study: an anchor test random groups design (Peterson et al, 1989). But they would take place in England, so that there would be no need to investigate the potential impact of curricular issues. All children taking part would already have been prepared for both the (parallel) tests concerned. Equating trials would require the assistance of sufficient schools to provide 1,000+ children in the required cohort for each test. Testing would be under secure conditions, managed by Test Development Agencies, just as pre-testing is at present. If test development and approval by government agencies could be guaranteed complete in time it might ideally take place shortly before children took their own operational national tests; to maximise motivation and provide useful experience for the children involved. Otherwise later in the summer term or even early autumn (using children just entering KS3) would be acceptable alternatives. Equating does not require samples to perform at an operational level, only that the full range of achievement is adequately represented. Equating would entail the spiral allocation of children to groups taking either the baseline test or the 'future' test being equated. Our Project's comparisons have shown that a suitable common measure of achievement is also desirable, to check the efficacy of random allocation and

[1] KS3 English would provide one partial exception to this, as a component will test literature and the set texts will change over the years, making it impossible to equate them directly. However it will be possible to equate the remaining component of KS3 English to the equivalent component in the reserve test, should this be used as the baseline instrument.

[2] By switching resources away from the current second pre-test, which would not need the large numbers of children currently involved if it lost its present equating role.

[3] The need for all agencies and government departments to take heed of this stricture cannot be over-emphasised. If changes to tests are required after equatings the equatings are rendered void. Without valid equating it is impossible to recognise any growth in schools' achievements in the year in question, though it should not affect subsequent years.

provide a basis for statistical adjustment in equating if necessary. Fortunately the children's operational national test scores would be ideal for this purpose, so additional testing would not be required. Schools for such equating trials would need to be contacted well ahead to obtain suitable 'samples'[4].

Renewing baseline instruments

The same baseline instrument would be retained until changes to the curriculum made it invalid, probably when the next renewal cycle took effect. The curricular changes involved then govern the action required. If curricular change is not too extensive it may be possible to equate the existing baseline instrument to its successor (and perhaps the first operational test) through direct experimental equatings like those described above. But if radical curricular change negated this, a conservative approach to the use of the other sources of information routinely available in setting national test thresholds should enable a successful transfer of existing standards to a new regime.

A logical basis for deciding test thresholds at the FLTSM

Given the type of assessments used in national tests in England, setting standards is inevitably a 'social and societal process' (Whetton, Twist & Sainsbury, 2000 - who provide an authoritative account of standard setting in this context), involving empirical and qualitative information which requires some interpretation. The integrity of the interpretations and the decision taking process is vital if the legitimacy of these assessments is to be maintained and without this they will have no real value. As England's national testing system has evolved, QCA and the agencies responsible for the development of these national tests have gradually improved and refined the procedures and information available to help to set equivalent national test standards from one year to the next. Recent annual Final Level Threshold Setting Meetings (FLTSMs) have drawn upon a variety of sources of evidence, including the views of teachers and senior markers, quasi-experimental statistical equatings and, latterly, distributional data in which to model potential outcomes. Most of the ingredients required to produce effective decision taking are thus already in place. Our suggestions for improvement try to bring a logical process to bear as the information available is considered - and to suggest ways in which some sources might be improved.

The starting point - National Sample Data

The first step should be to make better use of the National Sample Data (NSD). They have a valid voice in answering a vital question authoritatively - 'on the assumption that standards of achievement in schools have not changed, which threshold marks produce the same pattern of results as last year?' These data should be presented first, before any other information, to suggest thresholds based on the null hypothesis that children's work has neither improved nor declined.

[4] These need to span the ability range but need not be strictly random or otherwise 'representative', as equating trials do not need to estimate population means. If a previous operational test were used as the baseline, schools would require at least a full year's notice and would have to agree not to use the relevant test in preparation or practice.

They should be seen as the 'natural' recommendations, which *will* be followed *unless* other sources of information can demonstrate sound evidence that things have improved or worsened. The onus of proof should lie with those who argue that things have changed. Now that the national curriculum and testing system is mature we should not be too surprised if results do not improve every year. Their introduction may have given the educational system a boost but it will become increasingly difficult to achieve real change in the quality of learning. The likelihood is that frequent and/or substantial year to year shifts in results on a national scale will owe more to errors in standard-setting than anything else.

In the absence of sound evidence supporting an alternative course, giving primacy to the null hypothesis would in itself reduce the very real risk that disparities between the various recommendations might sometimes make it difficult to avoid inaccurate outcomes.

Thresholds based on the null hypothesis would provide credibility when the quality of other information might be less than convincing - at the points of transition between cycles for instance, when both equatings and judgemental information might be less reliable than at other times and when schools and children will be adapting to new demands. They provide a sound and inherently conservative basis for decisions.

But whilst National Sample Data may provide a starting point and a safe recommendation to fall back on, they cannot recognise the national gains (or losses) in learning which might be made by schools. For this most essential element in a national assessment system we must look elsewhere.

The importance of hard evidence from improved equating

To prove that schools have produced gains (or lost ground) in achievement, the highest priority must be given to improving the quality of the statistical equatings brought to the table. These are of paramount importance in providing hard evidence to convince the FLTSM that the null hypothesis should be set aside. They are the *only* source of *empirical* information which addresses the question - 'if children took both today's and yesterday's tests, what marks on today's version would be equivalent to the past thresholds?'

Without better evidence of this sort we will never be able to defend decisions implying that the quality of educational output has really changed for the better (or worse) should they be challenged in public. Detailed recommendations about how to achieve this are presented under the heading 'A baseline equating strategy' above.

As the best source of hard evidence, the statistically equated thresholds should be presented immediately after the NSD and this should be seen as leading the case for setting aside the thresholds initially recommended. Equating is a conceptually and statistically complex matter and at times several techniques may be employed, producing a variety of results. There should be good grounds for choosing one technique/result over another, or a

defensible basis for compromise, but most members of the FLTSM will not be equipped to appreciate the complexities involved. This is not the place to explore alternative statistical models. But the TDA should be required to make it very clear how much confidence it has in the equating data in question and the recommendations it is putting forward. The FLTSM can then share their confidence or doubts whilst deciding whether equatings justify setting aside the NSD's initial thresholds - should the two disagree.

Expressing markers' views

The social dynamics of the FLTSM could make it hard to resist a confident assertion by a Lead Chief Marker (LCM) that the quality of work was much improved this year. To reduce the risk of conflict:

- the mark ranges considered by scrutineers should be centred on the NSD threshold recomendations representing the null hypothesis.
- to discourage a gradual 'drift' of standards, archive materials for Script Scrutineers should not be taken from the previous year alone. If a cyclical approach is taken to curriculum renewal, extra efforts might be made to preserve high quality archive material from the initial year in each cycle as a mainstay.
- we should recognise that Script Scrutiny is a judgemental process, and thus has limited precision (Creswell,1996 & 2000). Markers should no longer be asked to recommend a single threshold mark and the outcome of the scrutiny should simply be graphical displays of the ratings made, perhaps also summarised in the form of a zone of uncertainty within which markers think the threshold likely to fall (e.g. 34 - 38).

Giving teachers a more effective voice

Teachers' recommended thresholds - arising from the use of 'Angoff' procedures (Angoff,1971; Morrison et al, 1994), are gathered and presented by TDA's. Their low profile perhaps stems from not unreasonable (Shepard, 1980; Jaeger,1989) lack of confidence in this approach, especially where results conflict with NSD or equating evidence - as is often the case. It might be better to include a small group of 'active' classroom teachers alongside the senior external markers (many of whom will not currently be teaching the subject/key stage concerned) within script scrutiny exercises. The LCM could present teachers' views alongside the markers'. Agreement or disagreement between the two groups would itself be of interest. If the suggestions above are taken up, teachers taking part will in this way be pre-focussed on the relevant range of marks and graphical display of individual judgements should provide adequate means to decide how much faith to place in them.

More defensible decision-making

Whilst involving only modest alterations to current practice, this restructuring would mean the logical process by which the decision is reached is central and apparent to all concerned and where the burden of proof is clearly assigned. It ought also to ensure that the year on year instability of standards sometimes encountered in the recent past is minimised and help to explain decisions which imply that the quality of schools' output has changed for better or worse: thus enhancing the legitimacy of national assessments.

Future monitoring of national test standards over time

Independent audit of equating data

If the suggestions made here for improving the equating arrangements for national tests were adopted, little purpose would be served by repeating the kind of experimental comparisons which formed the mainstay of this project. Such work would have been built-in to the test thresholds set.

However, it may provide some public re-assurance if it was understood that, after an extended period, independent audits of threshold setting decisions in each key stage / curriculum area were to be undertaken and their reports made public. Audits would consider equating evidence, NSD threshold recommendations and distributional data, and the other sources of information available to FLTSM. Files which would be suitable for audit at a later date would need to be produced annually, consisting largely of the FLTSM papers, with some additional details regarding equatings. It might be desirable to ask an auditing agency to observe the standard setting process each year and to participate in the construction of the files, to ensure that these prove fit for purpose.

Curriculum renewal cycles of about 5 years would again provide appropriate intervals between audits, with some flexibility to keep step being required should renewal be advanced or postponed.

Long term judgemental comparisons

Objective comparisons of standards over time intervals longer than each curriculum renewal cycle are almost certain to prove difficult. Change in the curricular and assessment regimes will be allied to changes in educational practice and social values; making it impractical to ask children to take tests from two eras. But we might consider judgemental comparisons, where we would at the very least learn more about the nature of changes in what is cherished. If it is thought that this might be wanted at some future date, the design of such a study should be considered now, so that appropriate samples of children's work can be archived explicitly for this role. The final year of any phase of curriculum renewal would appear to be a prime candidate for inclusion in such studies and it might be highly desirable to archive such samples of scripts from the 2002 national tests.

Though likely to detect only relatively substantial shifts in test standards, this strategy would provide some assurance of equivalence over the long term. If carried out after the final year of each cycle such studies might also inform the transitions between one curriculum / assessment phase and the next. Whether it would be feasible to try to mount such research as the current cycle concludes in 2002 would depend on the availability of sample scripts from the past.

Evidence from LEA standardised testing programmes

We have explained why LEA standardised testing data cannot, of itself, provide conclusive evidence relating to standards over time. But such data have provided valuable supplementary evidence in this project and it might be

appropriate to try to gather together information of this sort on a long term basis, and to analyse it in ways calculated to help inform QCA's management of the national testing programme.

Annual surveys of achievement standards

Another option would be the collection of independent data relating to movements in achievement through annual 'surveys' of achievement involving soundly based national samples of children. This could be undertaken in every curriculum area at each key stage or only in selected subjects. Alternative methods of achieving this are available and choice between them may depend on the views taken about the relative importance of uses to which such data might be put. For instance a sophisticated domain sampling approach like that of the Assessment of Performance Unit (Johnson, 1989; Johnson and Bell, 1985) could employ a variety of different assessments set to different children to estimate achievement over a wide curricular range. At the other extreme a single quite short test (either an existing standardised test or one tailor-made for the purpose) could provide a cruder but relatively inexpensive means of monitoring overall gains or losses in learning. It might even prove feasible to integrate this latter approach into the annual standard setting process. For instance, it might be possible to sample a sub-set of the children who are to be included in the National Sample. Then, if such tests were machine marked, it might just prove feasible to meet time schedules allowing consideration of evidence of year to year variations at the FLTSM; whilst the NSD's implications for threshold setting are under discussion. Alternatively, such data could be considered after standard-setting has been completed, with a view to informing the next year's decisions. The use of item based calibration models would allow partial replacement or supplementation of such tests at points of curricular renewal, which might greatly assist in transition years, when the curriculum and assessment regimes have been revised.

Beyond national tests

The contribution of national tests should not be under-rated. There can be little doubt that the most important of this Project's findings is the data which provide sound evidence that, since the advent of national tests, achievement levels in schools have in fact improved substantially in almost all curriculum areas/key stages investigated. The suggestion that in some instances variations in test standards might account for some (but by no means all) of the changes in national results must not be allowed distract attention from this. Given that the evidence from the past half century shows how hard it is for educational innovations to achieve such progress, this is a remarkable achievement, which should be celebrated widely. The introduction of the testing system allied to curriculum change, together with publication of results at school level, was intended to stiffen motivation to improve teaching and learning and its contribution may well have been significant.

But the system is now maturing and whilst it may continue to have a role to play for some time to come we should not assume that it will be needed in perpetuity. After all, as Aldrich (2000) has recently entertainingly pointed out, most initiatives have a limited shelf life and even the late nineteenth century payment by results system was eventually seen to handicap rather than help. Although payment by results may have lasted about forty years, the pace of change as we enter the 21st century is quicker. National testing in its current form is expensive, primarily because of the external marking of the tests, and the time may soon come when it is thought that these resources may make a better contribution elsewhere. Has this Project any evidence which might suggest how the purposes served by national testing could be achieved less expensively?

An interesting minor feature of the evidence from LEA standardised testing programmes was the indication that at both key stage 1 and 2, teacher assessments showed less sign of drifting standards than national tests in Reading/English or Mathematics. Teacher assessment appears in this light less unreliable than might have been assumed when the current national testing system was designed. Might it have a role to play in the future? It is not necessarily a question of either / or. Teacher assessments might provide the assessment of individual children, whilst shorter national tests (perhaps not unlike those suggested for annual surveys of achievement standards above) could be used to moderate differences between schools or even individual teachers. If such tests were largely automatically marked the costs would be comparatively modest. But they would still enable a national assessment system to monitor national progress effectively, and by providing a basis for assessments of equivalent standard across schools, could continue to be used to motivate further improvement. Wiliam (2001) recently advocated something similar and this idea is far from new. For instance Sweden applies a similar scheme to assessments marking the end of compulsory schooling (Wolf, 2000).

Their day may be some time off, but we should soon begin to think about more radical re-arrangements for national assessment.

7 References

Ahmed, A. & Pollitt, A. (1999) 'Curriculum demands and question difficulty', Paper to the International Association for Educational Assessment Conference, Bled, Slovenia.

Ahmed, A. & Pollitt, A. (2001) 'Improving the validity of contextualised questions', Paper at British Educational Research Association Conference, University of Leeds.

Aldrich, R. (2000) 'Educational standards in historical perspective', Proceedings of the British Academy, 102, 39-67.

Angoff, W.H. (1971) 'Scales, norms and equivalent scores', in R.L.Thorndike (Ed.) Educational Measurement, 2nd Edition, American Council on Education, Washington D.C.

Beard, R. (1998) 'National Literacy Strategy: Review of research and other related evidence', DfEE, London.

Bell, J.F., Bramley, T. & Raikes, N. (1998) 'Investigating A level mathematics standards over time', British Journal of Curriculum and Assessment, 8, 2, 7-11.

Bond, L. (1993) 'The effects of special preparation on measures of scholastic ability', In R. Linn (Ed.) Educational Measurement, 3rd Edition, Oryx Press for National Council on Measurement in Education & American Council on Education, Phoenix AZ.

Bramley, T., Bell, J.F. & Pollitt, A. (1998) 'Assessing changes in standards over time using Thurstone paired comparisons', Education Research and Perspectives, 25, 2, 1-24.

Brooks, G., Foxman, D. & Gorman, T. (1995) 'Standards in Literacy and Numeracy: 1948-1994', National Commission on Education Briefing, New Series, 7, London.

Brown, B. (1999) Reported in Pyke, N. '1999 reading test easier, says head', Times Educational Supplement, 29.10.99.

Brown, M., Taggart, B., McCallum, B. & Gipps, C. (1996) 'The impact of key stage 2 tests', Education 3 to 13, October 1996, 3-7.

Brown, M., McCallum, B., Taggart, B. & Gipps, C. (1997) 'The validity of national testing at age 11: the teacher's view', Assessment in Education, 4, 2, 271-293.

Christie, T. & Forrest, G.M. (1981) 'Standards at GCE A-level: 1963 and 1973', Schools Council Publications/Macmillan Education, London.

Clarke, C. (1997) ‘The impact of national curriculum statutory testing at key stages 1 and 2 on teaching and learning and the curriculum’, British Journal of Curriculum and Assessment, 7, 1, 12-18.

Coe, R. (1999) 'Changes in examination grades over time: is the same worth less?', Paper at British Education Research Association Conference, Brighton.

Colwill, I. (1997) ‘Intentions and perceptions: a review of the first year of monitoring of the school curriculum in England’, British Journal of Curriculum and Assessment, 7, 1, 33-37.

Cresswell, M. (1994) 'Aggregation and awarding methods for national curriculum assessments in England and Wales', Assessment in Education, 1, 45-61.

Cresswell, M. (1996) 'Defining, setting and maintaining standards in curriculum embedded examinations: judgemental and statistical approaches', In Goldstein, H. & Lewis, T. (Eds.), 'Assessment: Problems, Developments and Statistical Issues', Wiley, London.

Cresswell, M. (2000) 'The role of public examinations in defining and monitoring standards', Proceedings of the British Academy, 102, 69-120.

Daniels, S. & Stainton, R. (1994) ‘Revealing Results’, Primary Teaching Studies, Summer 1994, 39-42.

Davies, J. (1999b) 'Standards in mathematics in Year 6: what do key stage 2 national tests tell us?', Educational Psychology, 19, 1, 71-77.

Davies, J. & Brember, I. (1995) ‘The first and second mathematics standard assessment tasks at key stage 1: a comparison based on a five schools study’, Educational Review, 47, 1, 3-9.

DfEE (1997) 'From Targets to Action', Department for Education and Employment, London.

DfEE (2001a) 'Schools - building on success', Department for Education and Employment, London.

DfEE (2001b) 'National curriculum assessments of 7, 11 and 14 year olds, 2001', Department for Education and Employment, London.

Dickens, W. & Flynn, J. (2001) 'Heritability estimates versus large environmental effects: the IQ paradox resolved', Psychological Review, 108, 2, 346-369.

Earl, L., Levin, B., Leithwood, K., Fullan, M., Watson, N., with Torrance, N., Jantzi, D. & Mascall, B. (2001) 'Watching & Learning 2: OISE/UT evaluation of the implementation of the National Literacy and Numeracy Strategies, Ontario Institute for Studies in Education, Ontario.

Fitz-Gibbon, C. & Vincent, L. (1994) ‘Candidates’ performance in public examinations in Mathematics and Science’, School Curriculum and Assessment Authority, London.

Fox, B. (1998) ‘Improving Marks at Key Stage 2’, British Journal of Curriculum and Assessment, 7, 2, 7-19.

Fransella, F. & Bannister, D. (Eds) (1977) 'A manual for repertory grid technique', Academic Press, London.

Galton, M. (1998) 'Back to consulting the ORACLE', Times Educational Supplement, 3.7.99.

Gipps, C. & Murphy, P. (1994) ‘A Fair Test?’, Open University Press, Buckingham.

Goldstein, H. (1983) 'Measuring changes in educational attainment over time: problems and possibilities', Journal of Educational Measurement, 20, 4, 369-377.

Good, F. & Cresswell, M. (1988) ' Grading the GCSE', Secondary Examinations Council, London.

Greatorex, J. (2001) 'Making the grade - how question choice and type affect the development of grade descriptors', Educational Studies, 27, 4, 451-464.

Greatorex, J. (2002) 'Making Accounting examiners' tacit knowledge more explicit, developing grade descriptors for an Accounting A level', Research Papers in Education (in press).

Greatorex, J., Johnson, C. & Frame, K. (2001) 'Making the grade - developing grade profiles for Accounting using a discriminator model of performance', Westminster Studies in Education, 24, 2, 167-181.

Green, S. (2001) 'A study of the effects of content and structural support in writing tasks', Paper at 12th European Conference on Reading, Dublin, July 1-4.

Green, C., Hamnett, L. & Green, S. (2001) 'Children put national tests to the test', Education 3-13, 29, 3, 39-42.

Hilton, M. (2001) 'Are the Key Stage Two Reading Tests becoming easier each year?', Reading, Language and Literacy, April, 4-11.

Hurry, J. & Sylva, K. with Fox Lee, L. & Mirrelman, H. (1996) 'Standardised literacy tests in primary schools: their use and usefulness', School Curriculum and Assessment Authority, London.

Jaeger, R. (1989) 'Certification of student competence', in R.L.Linn (Ed.) Educational Measurement, 3rd Edition, Oryx Press for National Council on Measurement in Education & American Council on Education, Phoenix AZ.

Johnson, S. (1989) 'National Assessment: The APU Science Approach', HMSO, London.

Johnson, S. (1996) 'The contribution of large-scale assessment programmes to research on gender differences', Educational Research & Evaluation, 2, 1, 25-49.

Johnson, S. & Bell, J.F. (1985) 'Evaluating and predicting survey efficiency using generalizability theory', Journal of Educational Measurement, 22, 107-119.

Kelly, G.A. (1955) 'The Psychology of Personal Constructs, vols 1 & 2', Norton, New York.

Literacy Task Force (1997) 'A reading revolution: how we can teach every child to read well', The Literacy Task Force c/o University of London Institute of Education, London.

Massey, A.J. (1978) 'A model for screening against standards drift between years using information concerning pass rates in different types of centre in the maintained school sector: O level Mathematics and English Language examinations set by the Oxford Delegacy of School Examinations and the University of Cambridge Local Examinations Syndicate between 1969 and 1975. Test Development and Research Unit, Cambridge.

Massey, A.J. (1982) 'Assessing 16+ Chemistry: the exposure - mastery gap', Education in Chemistry, September, 143-145.

Massey, A.J. (1994) 'Standards are slippery', British Journal of Curriculum and Assessment, 5, 37-8.

Massey, A.J. (1995) 'Criterion-related test development and national test standards', Assessment in Education, 2, 2, 187-203.

Massey, A.J. (1997) 'The feasibility of equating national test standards in science between key stages 2 and 3 and from year to year', Educational Review, 49, 1, 29-45.

Massey, A.J. (1998) 'Equating standards in 1998 KS3 National Tests in England, Wales and Northern Ireland', Report to QCA, ACCAC & CCEA (unpublished), Research & Evaluation Division, UCLES, Cambridge.

Massey, A.J. & Elliott, G.L. (1996) 'Aspects of writing in 16+ English examinations between 1980 and 1994', Occasional Research Paper 1, UCLES, Cambridge.

Massey, A, Elliott, G & Ross, E. (1996) Season of birth, sex and success in GCSE English, mathematics and science: some long lasting effects from the early years? Research Papers in Education, 11, (2), 129-50.

Massey, A.J. & Newbould, C.A. (1978) 'Standards drift: a screening - UCLES Advanced Level 1969-1976, Test Development & Research Unit, Cambridge.

Morrison, H., Busch, J., & D'arcy, J. (1994) 'Setting reliable national curriculum standards: a guide to the Anghoff procedure', Assessment in Education, 1, 181-199.

Newbould, C.A. & Massey, A.J. (1979) 'Comparability using a common element', TDRU Occasional Publication 7, Cambridge.

Newton, P. (1997) 'Examining standards over time', Research Papers in Education, 12, 3, 227-48.

Newton, P. (2000) 'Enhancing the defensibility of procedures for maintaining national curriculum test standards', A Report for the Nuffield Assessment Seminar Group, University of London Institute of Education, NFER, Slough.

OFSTED & SCAA (1996) 'Standards in public examinations 1975 to 1995', School Curriculum and Assessment Authority, London.

Patrick, H. (1996) 'Comparing public examination standards over time', Paper to British Educational Research Association Conference.

Petersen, N.S., Kolen, M., & Hoover, H.D. (1989) 'Scaling, norming and equating', Chapter 6 in Educational Measurement (3rd edition), R.Linn (Ed), Oryx Press for National Council on Measurement in Education & American Council on Education, Phoenix AZ.

Plewis, I. & Veltman, M. (1996) 'Opportunity to learn Maths at Key Stage 1: changes in curriculum coverage 1984-1993', Research Papers in Education, 11, 2, 201-218.

Pollitt, A. & Ahmed, A. (1999) 'A new model of the question answering process', Paper to the International Association for Educational Assessment Conference, Bled, Slovenia.

Pollitt, A., Entwistle, N., Hutchison, C. & de Luca, C. (1985) 'What makes exam questions difficult', Scottish Academic Press, Edinburgh.

Preece, P. & Skinner, N. (1999) 'The national assessment in Science at Key Stage 3 in England and Wales and its impact on teaching and learning', Assessment in Education, 6, 1, 11-25.

Qualifications and Curriculum Authority (2001) 'Five yearly review of standards reports', QCA, London.

Reay, D. & William, D. (1999) 'I'll be nothing': structure, agency and the construction of identity through assessment', British Educational Research Journal, 25, 3, 343-354.

Rose, J. et al (1999) 'Weighing the baby', Report of the Independent Scrutiny Panel on the 1999 Key Stage 2 National Curriculum tests in English and Mathematics, Department for Education and Employment, London.

Sadler, D. (1987) 'Specifying and promulgating achievement standards', Oxford Review of Education, 13, 191-209.

Sainsbury, M. & Sizmur, S. (1998) 'Level descriptions in the National Curriculum: what kind of criterion-referencing is this?', Oxford Review of Education, 24, 2, 181-193.

Sharp, C., Hutchison, D. and Whetton, C. (1994) 'How do season of birth and length of schooling affect children's attainment at key stage 1?', Educational Research, 36, 2, 107-21.

Shepard, L. (1980) 'Standard setting issues and methods', Applied Psychological Measurement, 4, 447-467.

Sizmur, S. & Sainsbury, M. (1997) 'Criterion referencing and the meaning of national curriculum assessment', British Journal of Educational Studies, 45, 2, 123-140.

Whetton, C., Twist, E. & Sainsbury, M. (2000) 'National Tests and Target Setting: Maintaining Consistent Standards', Paper at American Educational Research Association Annual Meeting, New Orleans.

Wiliam, D. (2001) 'Level best? Levels of attainment in national curriculum assessment', Association of Teachers and Lecturers, London.

Willmot, A.S. (1997) 'CSE and GCE Grading Standards: the 1973 comparability study', Schools Council Research Study, Macmillan Education, London.

Willmott, A.S. (1980) 'Twelve years of examinations research; ETRU 1965-1977', Schools Council, London.

Wolf, A. (2000) 'A comparative perspective on educational standards', Proceedings of the British Academy, 102, 1-8.

Ziomek, R. & Svec, J. (1995) 'High school grades and achievement: evidence of grade inflation', American College Testing Program Research Report 95-3, ACT, Iowa City.