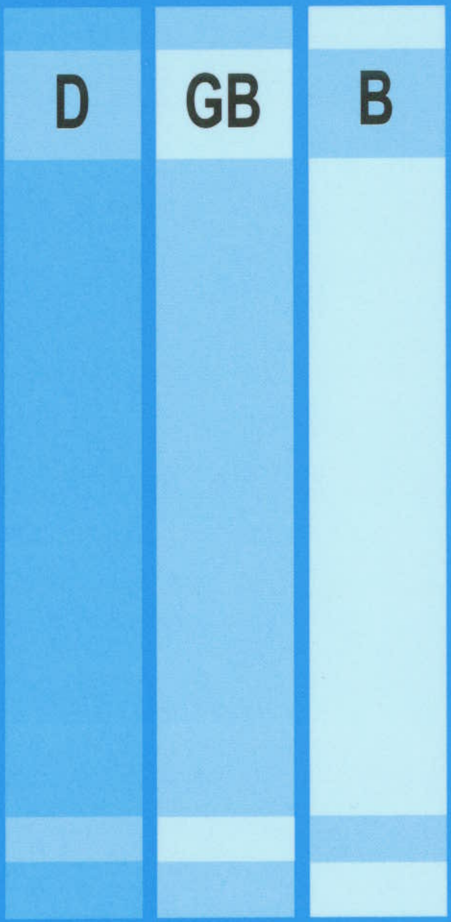
READING PERFORMANCE AT NINE



Greg Brooks
A.K. Pugh
Ian Schagen

nfer
The Open University

READING PERFORMANCE AT NINE

Greg Brooks
A. K. Pugh
Ian Schagen

nfer
The Open University

Published in July 1996
on behalf of the National Foundation for Educational Research
and the Open University
by NFER, The Mere, Upton Park, Slough, Berkshire SL1 2DQ

Registered Charity No. 313392
ISBN 0 7005 1431 7

CONTENTS

LIST OF TABLES

LIST OF FIGURES

PROJECT STAFF

Dr Greg Brooks and Tony Pugh were joint Directors of the project.

Dr Greg Brooks is a Senior Research Officer at the NFER.

Tony Pugh is a Staff Tutor, Faculty of Educational Studies, Open University Yorkshire Region, Leeds

Ian Schagen, Principal Research Officer and Head of Statistics at NFER, was the Project Statistician.

ACKNOWLEDGMENTS

We should like to record our appreciation to

- Mr Hugo de Burgh of DBF Television and Nottingham Trent University for suggesting the project
- the Esmée Fairbairn Charitable Trust for providing the principal funding for the project, and Channel 4 Television, the *Mail on Sunday* newspaper, the NFER and the Open University for additional support
- Dr Steve Sizmur and Dr Ralph Tabberer of NFER for comments on the report
- Mr Cres Fernandes of NFER for checking the statistical analyses
- the teachers who marked the tests
- Mrs Alison Bannerman, Miss Effie de Souza, Miss Mary Hargreaves, Mr Tim Wright and Dr Enver Carim for working on the publishing of the report
- Professor Neville Postlethwaite, Dr Andreas Schleicher, Dr Hans Pelgrum, Dr Rien Steen and others who were involved with the 1991 IEA study for much help, especially for providing copies of tests and questionnaires, statistical advice and large quantities of statistical information

and particularly to

- all the schools and pupils in the England and Wales who took part: without their cooperation none of this project would have been possible.

CHAPTER 1: BACKGROUND, AIMS AND KEY FINDINGS

1.1 Background

In March 1996, the National Foundation for Educational Research (NFER) and the Open University jointly carried out a survey of reading attainment in Year 4 classes (pupils aged 9) in England and Wales, and this is the report of that survey. The samples of schools and pupils were nationally representative. Most of the pupils involved had also been tested a year earlier, in Year 3 (when aged 8 on average).

The tests used in 1996 were:

- the test used in 1991 in a survey of 9-year-olds in 27 other countries (Elley, 1992). That survey was conducted by the International Association for the Evaluation of Educational Achievement (IEA). The data from this test in England and Wales in 1996 were analysed by exactly the same method as that used by IEA in 1991;
- level B of the (British) *Reading Ability Series* (Kispal *et al.*, 1989).

1.2 The purposes of the 1996 survey

The main purpose of the 1996 survey was to provide reliable evidence about

- international comparisons between England and Wales in 1996 and the 27 countries which had taken part in the 1991 IEA study; and
- the progress made by pupils in England and Wales between age 8 (Year 3) and age 9 (Year 4).

Subsidiary purposes were:

- to compare the *Reading Ability Series* level B results with those of the test's standardisation in 1987; and
- to compare performance on the IEA and *Reading Ability Series* tests.

1.3 Key findings

The key findings of the 1996 survey were that

- **the average score on the international test would have put England and Wales close to the overall average in the 1991 study, within a group of 13 countries whose average scores did not differ significantly;**
- **the average score for England and Wales on the international test was lowered by (among other factors) a 'long tail' of pupils who achieved scores well below the average; and**
- **the pupils tested in both 1995 and 1996 appeared to have made slower progress, on average, in the intervening 12 months than children did in 1987.**

It was also concluded that the international test was less suitable for pupils in England and Wales than the British test.

1.4 The structure of this report

The way the survey was carried out is described in outline in chapter 2 and in full in Appendix A. The results are presented in chapter 3, and conclusions are stated in chapter 4.

CHAPTER 2:
AN OUTLINE OF HOW THE SURVEY WAS CARRIED OUT

The first two sections of this chapter state the context of the 1996 survey, and the rest of the chapter gives a brief description of how it was carried out. A full description is given in Appendix A.

2.1 The 1991 IEA reading study, and Britain's non-participation

In the late 1980s, the IEA decided to mount an international study of reading achievement in 1991. The study was to focus on students of two ages (9 and 14), and all member countries were invited to participate. With the agreement of the (then) Department of Education and Science, the NFER became involved as the organising institution for England and Wales, and participated in the early planning and a pilot survey.

However, the researchers involved at NFER became increasingly dissatisfied with the tests which were being developed for use in the main study. They consisted almost entirely of multiple-choice items, and focused almost entirely on literal comprehension – in short, they were felt to represent an outmoded and inadequate model of the reading process (see *The Times Educational Supplement*, 16 October 1992, p.14). England and Wales accordingly withdrew from the study, which however went ahead in 32 other countries; 27 participated at age 9 and 31 at age 14 (Elley, 1992).

2.2 The 1996 survey in England and Wales

However, when, in 1995, an opportunity arose to mount a partial England and Wales replication of the 1991 IEA study, and to contextualise it by parallel use of a British test, the NFER decided to take the opportunity. The survey took place in March 1996 and was confined to age 9 (pupils in Year 4). The 1991 IEA test for age 9 was used (slightly modified – see Appendices A and C). In order to compensate for the deficiencies in the IEA test, for comparison and as part of the contextualisation, a British test was included in the study. This test was level B of the *Reading Ability Series* (Kispal *et al.*, 1989) – referred to at most points in this report as *RAS*.

An additional motive for mounting a reading survey in Year 4 in England and Wales in 1996 was that a survey had been carried out in Year 3 in 1995, using *RAS* level A. That survey had included a freshly-drawn, nationally representative sample of schools and pupils. In order to introduce a longitudinal element into the 1996 survey, that sample of schools was approached again, with the intention that the 1996 sample should consist largely of pupils who had already been tested in 1995; and the purpose of using the next level of the same test series was to maximise the statistical reliability of the longitudinal comparison.

The 1995 Year 3 survey was in turn the third in a series of such surveys. It would therefore also be possible to look back at the results of the two previous Year 3 surveys (1987, 1991).

2.3 The IEA test

The results reported for this two-part international test in chapter 3 were based on 60 items. Four of these were simple 'supply' (open-ended) items, and all the rest were multiple-choice (each with four options). The questions were based on 15 mostly very short texts. The texts were classified by the original devisers into three 'Domains': Narrative, Expository (factual), and Documents. The last category was defined as follows:

> *Structured information displays presented in the form of charts, tables, maps, graphs, lists or sets of instructions. These materials were organised in such a way that students had to search, locate and process selected facts rather than read every word of continuous text. (Elley, 1992, p.4)*

The numbers of texts and items in the IEA test are shown in Table 2.1.

Table 2.1: Numbers of texts and items in IEA test, by Domain and overall

	Number of texts	Number of items		
Domain		Part 1	Part 2	Total
Narrative	4	10	10	20
Expository	5	8	11	19
Documents	6	6	15*	21
Overall	15	24	36	60

* *Four of the items in this category were of 'supply' (open-ended) type; all others in the test were multiple-choice*

2.4 *Reading Ability Series* level B

A fuller description of this test in given in Appendix B. It consists of one Narrative text and one Expository text, and the numbers of items are as shown in Table 2.2.

Table 2.2: Numbers and types of items in *Reading Ability Series* level B

	Number and type of item		
	Multiple-choice	Open-ended	Total
Part 1 (all Narrative)	13	4	17
Part 2 (all Expository)	10	8	18
Total	23	12	35

2.5 Comparison of the two tests

Both tests were described as tests of reading ability, yet from Tables 2.1 and 2.2 it is clear that they differed in important ways. The proportion of multiple-choice items was 93 per cent in the IEA test, 66 per cent in *RAS* level B. The IEA test contained 'Documents' texts in addition to Narrative and Expository; *RAS* level B did not.

The level of comprehension targeted by the items differed even more significantly. The volume on the 1991 results in the United States (Binkley and Rust, 1994) contains a comparison of the comprehension levels of the items in the IEA test and an indigenous test. Items were classified as focusing on the test-taker's personal response, on 'critical stance' (understanding the author's intention), on 'initial understanding', or on 'developing an interpretation' – the last two categories being subdivisions of literal comprehension. The resulting comparison is shown in Table 2.3, with *RAS* level B also analysed on the same basis.

The most important difference revealed here was that the IEA test contained no items testing comprehension above the literal, whereas about a third of those in *RAS* level B targeted this higher level.

The graphic impression made by the two tests was also quite different (see Appendices B and C), mainly because of a difference in the quality of printing. This gave the IEA test a somewhat out-of-date appearance.

Table 2.3: Comprehension level of items in IEA, US and British tests

Test:	NAEP	IEA		*RAS* level B	
	(%)	N	(%)	N	(%)
Comprehension level:					
initial understanding	(17%)	2	(3%)	15	(43%)
developing an interpretation	(17%)	58	(97%)	9	(26%)
personal response	(33%)	0	(0%)	11	(31%)
critical stance	(33%)	0	(0%)	0	(0%)

Key; N = number of items
NAEP = (United States) National Assessment of Educational Progress
RAS = *Reading Ability Series*

Note: Information on N *(number of items) was not available for the NAEP test.*
Source of NAEP and IEA data: Binkley and Rust (1994, p.179), modified to take account of items deleted from international analyses of the IEA test.

The differences between the two tests can be summed up by saying that ***Reading Ability Series* level B was a more suitable test for pupils of this age in England and Wales than the IEA test.**

2.6 The achieved sample

In all, 58 schools took part in the 1996 survey; they constituted a nationally representative sample of schools in England and Wales containing Year 4 pupils.

The IEA test was taken by 1,817 pupils, and *RAS* level B by 1,803 of the same group. Within that total there were 1,504 pupils who had taken *RAS* level A in 1995. The average age of the pupils was 9 years 0 months.

The numbers of pupils for whom results are reported in chapter 3 are not always the same as those just given; this is the result of missing information on individual pupils.

CHAPTER 3:
THE RESULTS

In this chapter, the first three sections state, respectively, the main results for the IEA test, the limitations of those results, and comparisons, within those limitations, between the results for England and Wales in 1996 and the 1991 IEA study. Section 4 gives the results for the *RAS* test. The remaining sections of the chapter present correlations between the two tests, and then findings on differences in performance between boys and girls, and between pupils receiving and not receiving free school meals. There were so few pupils in the sample for whom English was not the first language (fewer than 40) that statistics are not reported for this variable.

Wherever a result is described as statistically significant, it was so at the 5 per cent level (p<0.05) or better.

3.1 The IEA test results for England and Wales in 1996

The proportions of pupils in England and Wales getting individual items right on this test ranged from 22 per cent to 92 per cent. The average raw scores on the two parts of the test and overall are shown in Table 3.1.

Table 3.1: Average raw scores on IEA test

	Average raw score		%
IEA test part1	17.1	(out of 24)	71.3
2	20.5	(out of 36)	56.9
total	37.6	(out of 60)	62.7

In order to compare the England and Wales results with those of the 1991 study, the raw scores were converted into Rasch ability scores (see Appendix A, section 9), following exactly the procedure used by the IEA. It was recognised that there are difficulties with the use of this analysis method. However, its use was necessary to allow comparison with the IEA results. The Rasch results for the three Domains and overall are shown in the second column of Table 3.2.

Table 3.2: IEA test results for England and Wales in 1996

Domain	Average Rasch score (uncorrected)	(standard deviation)	Average Rasch score (corrected for age)
Narrative	503	(103)	514
Expository	498	(96)	509
Documents	487	(96)	498
Overall	496	(86)	507

Number of pupils: 1,817

Average international Rasch score: 500

Average international standard deviation: 100

On corrections for age, see the following section

The uncorrected average Rasch scores, for each of the Domains and overall, were therefore very close to the international average.

3.2 Limitations on the IEA test results

However, before the IEA test results for England and Wales can be compared in detail with those for the other countries involved in 1991, a number of factors have to be borne in mind. This section contains discussions of three such factors: a possible ceiling effect, possible lower representativeness of the pupil samples in other countries, and the low average age of the pupils in the England and Wales sample.

Possible ceiling effect

There was a difference in testing procedure between this survey and the IEA study: in that study, all pupils took both parts of the IEA test, whereas in this survey each pupil took only one part, and the two parts were taken by equivalent half-samples. The effect of this is described in section A.10. Briefly, because the overall scores for the Domains, and thence for the test as a whole, were calculated from the scores of individual pupils on the separate Domains, in this survey the ranges of scores for the Domains were shorter (in some cases, much shorter) than in the full test. Therefore, some parts of the test showed a ceiling effect (a bunching of scores at the upper end), and some pupils will have been unable to show the full achievement of which they were capable. This will have lowered the average scores to an extent which is unquantifiable. However, it seems likely that the England and Wales result would be broadly similar to that stated below.

Representativeness of pupil samples in different countries

There are two relevant aspects of this factor. First, in England and Wales, only about 1.5 per cent of pupils are in special schools; in many other countries the proportion is higher (for example, about 5 per cent in the Netherlands). This means that there is a higher proportion of pupils with special educational needs in mainstream schools in England and Wales than in other countries. This may in turn have contributed to the 'long tail' of low scores to be discussed below, and also have depressed the overall averages for England and Wales somewhat.

Secondly, many other countries (France and the United States, for example) operate 'grade-based promotional systems'; that is, pupils whose achievements are particularly low for their age may be made to repeat a year. Since such pupils are then not in the same school year as the majority of those to be tested for IEA purposes, they do not form part of the eligible cohort from which samples are drawn. This might have the effect of *raising* the average scores of countries which operate 'grade-based promotion'. Since repeating is almost entirely absent from the British system, no such effect would apply to the England and Wales results.

Low average age of the England and Wales sample

The average age of the pupils tested in England and Wales was 9 years 0 months. Of the 27 countries in the 1991 study, only one (Canada/British Columbia) tested a sample whose average age was lower (8.9 years, or about 8 years 10.8 months). All the rest tested samples whose average age was higher, in some cases considerably so; the highest was Indonesia, with an average age of 10.8 years (about 10 years 9.6 months). The international average in 1991 was 9.7 years, or about 9 years 8.4 months. Hence the England and Wales sample was somewhat younger than than those from nearly all other countries.

From the discussion of corrections for age in the official international report on the 1991 study (Elley, 1992, Appendix E, especially p.107) it was estimated that a reasonable correction for the low average age of the England and Wales sample would be to increase each of the average Rasch scores in the second column of Table 3.2 by 11 points. This was confirmed by a regression analysis of age on average score for the 12 'month cohorts' of the England and Wales sample; this suggested a rise of 1.36 Rasch score points per month. For 0.7 of a year, or 8.4 months, this gave an average age-correction of 11.42 points. Rounded to 11 and added to the uncorrected Rasch scores, this gave the scores corrected for age shown in the right-hand column of Table 3.2.

3.3 Comparisons between England and Wales and 27 other countries on the IEA test

The average scores for England and Wales, uncorrected for age, are shown again in Table 3.3, in which they are compared with the uncorrected scores for the 27 countries which took part in 1991.

On the basis of the **uncorrected** overall score, England and Wales would have **been placed between Slovenia and Netherlands**. However, this placing must not be taken as absolute, for all the reasons given in the previous section and for an important additional reason. **The uncorrected average score for England and Wales in 1996 did not differ significantly from those of 13 countries which took part in 1991**; these are the countries in the shaded section of Table 3.3. Only the nine countries above the shaded section had average scores which were significantly higher than that for England and Wales; and only the five countries below the shaded section had average scores which were significantly lower.

The imprecision of the exact position of England and Wales within the middle group of 13 other countries can be illustrated by applying corrections for the one factor for which corrections can be precisely calculated, namely the average ages of the samples of pupils. As shown in the previous section, the England and Wales sample was relatively young, and should be given an increase of 11 points, for each of the three Domains and overall. Average scores corrected for age for the other 27 countries are given in Elley (1992, Appendix E, pp.107-111), and from that information Table 3.4 was compiled.

On the basis of the overall score **corrected for age**, therefore, England and Wales would have **been placed between West Germany on one side and French-speaking Belgium and Hungary on the other**, several places higher than on the basis of uncorrected scores, but still within the middle group of countries.

A prominent feature of the England and Wales results on the IEA test was a 'long tail'. In the top three-quarters of the distribution of Rasch scores, the pattern for England and Wales was very similar to that for many other countries. However, this was not true of the lowest quarter of the distribution, where that for England and Wales extended substantially further down the Rasch scale than the distributions for most other countries. This can be explained from the average scores for pupils at various percentile points, as shown in Table 3.5. The values shown are those for the Narrative Domain only, to facilitate comparison with Figure 3.1 in Elley (1992, p.19).

Table 3.3: Average uncorrected Rasch scores (with standard errors of sampling and standard deviations) for all Domains and overall, arranged in order of overall achievement, for the IEA 1991 age 9 study, with England and Wales scores for 1996

Country	Grade tested	Mean Age (in years)	OVERALL		NARRATIVE		EXPOSITORY		DOCUMENTS	
			Mean (s.e.)	SD	Mean (s.e.)	SD	Mean (s.e.)		Mean (s.e.)	SD
Finland	3	9.7	569 (3.4)	70	568 (3.0)	83	569 (3.1)	81	569 (4.0)	88
United States	4	10.0	547 (2.8)	74	553 (3.1)	96	538 (2.6)	80	550 (2.7)	81
Sweden	3	9.8	539 (2.8)	94	536 (2.6)	100	542 (2.7)	112	539 (3.2)	106
France	4	10.1	531 (4.0)	74	532 (4.1)	93	533 (4.1)	84	527 (3.9)	81
Italy	4	9.9	529 (4.3)	80	533 (4.0)	88	538 (4.0)	95	517 (4.9)	92
New Zealand	5	10.0	528 (3.3)	86	534 (3.5)	102	531 (3.1)	93	521 (3.3)	92
Norway	3	9.8	524 (2.6)	91	525 (2.8)	102	528 (2.3)	103	519 (2.8)	101
Iceland†	3	9.8	518 (0.0)	85	518 (0.0)	95	517 (0.0)	101	519 (0.0)	91
Singapore	3	9.3	515 (1.0)	72	521 (1.1)	91	519 (1.0)	75	504 (1.0)	78
Hong Kong	4	10.0	517 (3.9)	71	494 (4.1)	87	503 (3.4)	72	554 (4.2)	89
Switzerland	3	9.7	511 (2.7)	83	506 (2.6)	92	507 (2.7)	100	522 (2.8)	96
Ireland	4	9.3	509 (3.6)	79	518 (3.7)	94	514 (3.2)	89	495 (3.8)	84
Belgium/Fr	4	9.8	507 (3.2)	77	510 (3.3)	92	505 (2.8)	85	506 (3.5)	88
Greece	4	9.3	504 (3.7)	75	514 (3.8)	88	511 (3.6)	85	488 (3.8)	85
Spain	4	10.0	504 (2.5)	78	497 (2.4)	86	505 (2.3)	92	509 (2.7)	89
Germany/W	3	9.4	503 (3.0)	84	491 (2.8)	93	497 (2.9)	104	520 (3.2)	94
Canada/BC	3	8.9	500 (3.0)	80	502 (3.5)	96	499 (2.7)	94	500 (2.8)	86
Germany/E	3	9.5	499 (4.3)	84	482 (4.2)	93	493 (3.6)	103	522 (5.0)	96
Hungary	3	9.3	499 (3.1)	78	496 (2.9)	80	493 (3.1)	101	509 (3.5)	89
Slovenia	3	9.7	498 (2.6)	78	502 (2.7)	94	489 (2.5)	93	503 (2.5)	82
England & Wales (1996)	4	9.0	496 (5.3)	86	503 (6.3)	103	498 (5.9)	96	487 (5.8)	96
Netherlands	3	9.2	485 (3.6)	73	494 (3.3)	85	480 (3.4)	87	481 (3.9)	82
Cyprus	4	9.8	481 (2.3)	77	492 (2.4)	92	475 (2.3)	91	476 (2.1)	81
Portugal	4	10.4	478 (3.6)	74	483 (3.3)	81	480 (3.0)	84	471 (4.5)	92
Denmark	3	9.8	475 (3.5)	111	463 (3.4)	119	467 (3.5)	127	496 (3.6)	125
Trinidad/ Tobago	4	9.6	451 (3.4)	79	455 (3.6)	91	458 (3.4)	93	440 (3.3)	82
Indonesia	4	10.8	394 (3.0)	59	402 (2.8)	66	411 (3.2)	77	369 (3.0)	66
Venezuela	4	10.1	383 (3.4)	74	378 (3.2)	86	396 (3.3)	91	374 (3.7)	84

†Iceland tested all students, therefore no standard error was calculated.

s.e. = 1 standard error of sampling

Sources: – for England and Wales, the present study

– for all other data, Elley (1992), Table 3.1, p.14, slightly modified (Singapore moved above Hong Kong, and shading added to show countries whose average scores were not significantly different from that for England and Wales)

Table 3.4: Average overall Rasch scores, corrected for age and arranged in order of overall achievement, for the IEA 1991 age 9 study, with England and Wales score for 1996

Country	Grade tested	Mean Age (in years)	OVERALL	
			Mean (s.e.)	SD
Finland	3	9.7	570 (3.4)	70
United States	4	10.0	543 (2.8)	74
Sweden	3	9.8	539 (2.8)	94
Italy	4	9.9	528 (4.3)	80
France	4	10.1	526 (4.0)	74
New Zealand	5	10.0	524 (3.3)	86
Norway	3	9.8	524 (2.6)	91
Singapore	3	9.3	522 (1.0)	72
Iceland†	3	9.8	518 (0.0)	85
Ireland	4	9.3	516 (3.6)	79
Canada/BC	3	8.9	514 (3.0)	80
Hong Kong	4	10.0	514 (3.9)	71
Switzerland	3	9.7	512 (2.7)	83
Greece	4	9.3	511 (3.7)	75
Germany/W	3	9.4	508 (3.0)	84
England & Wales (1996)	4	9.0	507 (5.3)	86
Belgium/Fr	4	9.8	506 (3.2)	77
Hungary	3	9.3	506 (3.1)	78
Germany/E	3	9.5	504 (4.3)	84
Spain	4	10.0	500 (2.5)	78
Slovenia	3	9.7	499 (2.6)	78
Netherlands	3	9.2	494 (3.6)	73
Cyprus	4	9.8	481 (2.3)	77
Denmark	3	9.8	475 (3.5)	111
Portugal	4	10.4	468 (3.6)	74
Trinidad/ Tobago	4	9.6	454 (3.4)	79
Indonesia	4	10.8	378 (3.0)	59
Venezuela	4	10.1	368 (3.4)	74

†Iceland tested all students, therefore no standard error was calculated.

s.e. = 1 standard error of sampling

Sources:
- for England and Wales, the present study
- for all other data, Elley (1992), Table E.1, p.108, re-arranged in order of overall achievement

Table 3.5: Selected uncorrected percentile point values for England and Wales, 1996, Narrative Domain only

Percentile	Average uncorrected Rasch score
5th	303
25th	436
75th	582
95th	629

The England and Wales scores for the 95th and 25th percentiles were close to those of many countries in the middle range on the Narrative Domain; and the score for the 75th percentile was *higher* than that of all but four or five countries. But the score for the 5th percentile appeared substantially *lower* than that of all but three other countries.

In interpreting this 'long tail', both the absence of pupils repeating a year and the higher proportion of children with special educational needs in mainstream schools in England and Wales need to be borne in mind. On the other hand, the superior performance of England and Wales pupils whose scores fell around the 75th percentile would partly compensate for those factors; and it seems unlikely that the whole of the long tail could be accounted for by them. At least part of the long tail in the results for England and Wales therefore truly reflected lower attainment.

3.4 The *Reading Ability Series* results

The results for *RAS* level B in 1987 and 1996 are shown in Table 3.6.

Table 3.6: *Reading Ability Series* level B results, 1987 and 1996

	1987	1996
Average raw score (out of 35)	22.9 (65.4%)	22.0 (62.9%)
Average standardised score	100.0	98.8
(standard deviation)	(15.0) (15.2)	
Number of pupils	2126	1776

The difference between the two average standardised scores was **not** statistically significant. This result showed that **the pupils tested in 1996 performed on average at about the same level as those in the 1987 standardisation sample.**

This suggests that the average reading level for this age group had neither risen nor fallen, relative to the 1987 standardisation.

The results for *RAS* level B in 1996 and for level A in 1995 are shown in Table 3.7. This comparison is based only on the pupils who were tested in both years.

Table 3.7: *Reading Ability Series* results for level A (1995) and level B (1996)

	Level A **1995**	**Level B** **1996**
Average raw score	14.4 (57.6%)	22.4 (64.0%)
Maximum score on test	25	35
Average standardised score	101.8	99.4
(standard deviation)	(14.5)	(15.0)
Average Series Scale score	24.8	30.8
Number of pupils	1504	1504
	(N.B. same pupils in both years)	

Correlations of raw scores between levels A and B:

- in 1987 standardisation (taken 1 week apart): 0.81 (Kispal *et al.*, 1989, p.69; Number of pupils = 709)
- in this study (taken 1 year apart): 0.74

All tests have a degree of imprecision, and when different tests are used and compared the imprecision increases, additively. The correlation of 0.81 between levels A and B in 1987 showed that, just a week apart, a proportion of children changed their position in the rank order. The correlation of 0.74 in this study showed that a higher proportion had changed their positions after a year. Some will have made excellent progress, others much less; this needs to be borne in mind when considering average progress.

Between 1995 and 1996 the average **standardised** score of the pupils in this study who were tested in both years had **fallen**, by 2.4 standardised score points, and this fall was statistically significant.

This finding can be illuminated by using a related source of information in the *Teachers' Handbook* to the series (Kispal *et al.*, 1989). Table 2 on page 75 of the *Handbook* provides conversions of raw scores to what are described as 'Series Scale Scores'; through these, performances on adjacent levels of the series can be compared, and therefore progress between adjacent levels can be estimated. According to Table 5 on page 16 of the *Handbook*, the average gain in Series Scale scores between levels A and B in 1987 was 8.7 Series Scale points, which represents a substantial amount of progress. But, as shown in Table 3.7, the average gain in these scores for the pupils in this study was 6.0 points.

If these pupils had made progress in reading between March 1995 and March 1996 equivalent to the difference between the separate samples who took the two levels in 1987, then

- the average standardised score in 1996 of the pupils in this study who were tested in both years would have been statistically indistinguishable from their 1995 average score; instead, their 1996 average score was significantly lower; and
- their average gain in Series Scale Scores would have been closer to 8.7 than to 6.0.

These two statements are different ways of stating the same finding, and appeared to show that **the pupils tested in both 1995 and 1996 had made slower progress, on average, in the intervening year than children did in 1987**.

No factors that might have contributed to this finding could be deduced from this study, since it had not been designed to investigate such factors; however, the implications are discussed in chapter 4.

3.5 Correlations between the two tests

Correlations were calculated between overall scores on the two tests taken in 1996, and for comparison also between the domains of the IEA test. The results are shown in Table 3.8.

Table 3.8: Correlations within and between IEA and *Reading Ability Series* tests

		(1)	(2)	(3)	(4)	(5)	(6)
	IEA test						
(1)	– overall	1.00	0.88	0.88	0.86	0.78	0.75
(2)	– Narrative		1.00	0.68	0.62	0.70	0.68
(3)	– Expository			1.00	0.65	0.66	0.64
(4)	– Documents				1.00	0.67	0.64
	***RAS* LEVEL B**						
(5)	– raw score					1.00	0.97
(6)	– standardised score						1.00

The correlations within *RAS* level B and between the IEA test and each of its Domains were all satisfactory. Those among the IEA Domains, and between those and the IEA test overall on the one hand and *RAS* level B on the other, were all moderate. The most likely explanation of this is that the two tests as a whole, and the Domains within the IEA test, though all testing reading ability, were testing overlapping but partly complementary aspects of that general ability.

3.6 Differences in performance between boys and girls

The results for boys and girls on both tests are shown in Table 3.9.

Table 3.9: Results for boys and girls

Test	Average Rasch score uncorrected	 corrected for age	(standard deviation)	Number of pupils
IEA				
– boys	487	498	(90)	918
– girls	509	520	(78)	785
***RAS* LEVEL B**				
	Average standardised score			
– boys	96.9		(15.0)	932
– girls	101.3		(15.1)	823

The differences on both tests were significant. The superior performance of girls was consistent with that in many national reading surveys in Britain.

3.7 Pupils receiving and not receiving free school meals

The results on both tests for pupils receiving and not receiving free school meals* are shown in Table 3.10.

The differences on both tests were significant. The superior performance of pupils not receiving free school meals was again consistent with that in many national reading surveys in Britain.

Within tables 3.9 and 3.10 there is a further noteworthy feature. On the IEA test, not only were the average scores for boys and for pupils receiving free school meals substantially lower than those for girls and for pupils not receiving free school meals; the standard deviations for boys and for pupils receiving free

* *In England and Wales, almost all state schools provide a midday meals service. Children from low-income families (where the criterion of 'low income' is nationally defined) receive these meals free, and the cost is borne by the Local Education Authority.*

Table 3.10: Results for pupils receiving and not receiving free school meals

	Average Rasch score		(standard	Number of
Test	**uncorrected**	**corrected for age**	**deviation)**	**pupils**
IEA test				
– pupils receiving free meals	451	462	(89)	237
– pupils not receiving free meals	508	519	(81)	1146
***RAS* LEVEL B**				
	Average standardised score			
– pupils receiving free meals	91.3		(15.0)	251
– pupils not receiving free meals	100.8		(14.7)	1179

school meals were also larger. One implication of these findings is that among the pupils consisting the 'long tail' on the IEA test there were significantly more boys than girls, and significantly more pupils receiving than not receiving free school meals.

CHAPTER 4:
WHAT DID THE RESULTS MEAN?

Three of the results stand out:

- the average performance of the pupils involved in this survey was in the average group by international standards, whether uncorrected or corrected for age;
- the average score for England and Wales on the international test was lowered by the 'long tail'; and
- those tested in both 1995 and 1996 had made slower progress, on average, in the intervening year than children did in 1987.

4.1 The international comparison

From the international result it is impossible to deduce any trend over time. The only prior international survey of *reading* performance (before the 1991 study) was that carried out by the IEA at ages 9 and 14 in the early 1970s. England and Wales did take part on that occasion (as did Scotland, separately), and the England and Wales rank was fourth out of 14 countries (Thorndike, 1973, Table 8.1, p.124). But the tests then and in 1991 were different; the large gap in time makes any comparisons tenuous; and there were only seven countries in common between the two studies.

The only other international literacy survey at school level was the IEA study of *writing* carried out in 1983 (Gorman *et al.*, 1988); but no data from which a rank order of countries could be compiled were reported from that study.

The international result from the present study therefore has to be interpreted as a one-off event.

It seems unlikely that the IEA tests were too demanding for pupils in England and Wales; on the contrary, the questions were almost all pitched at a level of literal interpretation of the texts presented. From inspection of the six items which were excluded from the international statistical analyses (see Appendix A), it was clear that most concerned author's intention or the theme of a text or paragraph – that is, they were the higher-level questions. Therefore the 60 items which were analysed required hardly any comprehension beyond the literal, and the test should not have been either too difficult or too easy for 9-year-olds in England and Wales. In any case, literal comprehension is necessary as a basis for higher-order comprehension.

However, a factor which might be thought to have had some influence was familiarity with being tested. For the 1,504 pupils in this study who had been tested in 1995, the 1995 survey may well have been their only previous experience of taking a test under formal 'examination' conditions; and the 314 who had not been involved in 1995 may never have taken such a test before at all. Some of these pupils, both those tested before and those not – but not all of them because of industrial action at the time – would have taken Key Stage 1 national tests in Spring/Summer 1994; but those tests were administered in quite a different fashion. By comparison, the United States pupils tested in 1991 (whose average age was 10 years 0 months, a full year older than those tested in England and Wales in 1996) may have been much more 'test-wise'. But even if this was true for pupils in the United States (which was second in the overall 'league table'), it is not clear that it would be true for many of the other countries in the 1991 study.

More important would be the factors to which attention was drawn in chapter 3: the England and Wales results may have been lowered by a ceiling effect and by the presence in the sample of a higher proportion of children with special educational needs, and were certainly lowered by the younger average age. Other countries' scores may have been raised by the exclusion of pupils retained in lower years. Yet even when precise allowance was made for age, and account taken of the other factors just mentioned, it seemed clear that England and Wales would remain in the middle group of countries.

4.2 The 'long tail'

An important finding in this study, as in other international comparisons, was the existence of a 'long tail' in the results. Performance of lower ability pupils in England and Wales tails off drastically and tends to lower the average score in international comparisons.

The long tail in the scores on the IEA reading test in 1996 was consistent with

- similar distributions in mathematics and science in both national and international surveys (see the 'Numeracy' section of Brooks *et al.*, 1995);
- evidence that between a sixth and an eighth of the adult population of Britain has problems with basic literacy (see Ekinsmyth and Bynner, 1994, and the 'Literacy' section of Brooks *et al.*, 1995), and that this pattern has persisted for over 60 years (Adult Literacy and Basic Skills Unit, 1995).

Though this may be the major factor underlying the international result, that result requires national factors for its interpretation.

4.3 National factors

In this connection, the slow progress on the British test needs to be taken into account. The pupils who were tested in both years were clearly **not** a group whose attainment was inherently lower than that of other cohorts. The average standardised score in 1995 of those who took the tests in both years was actually slightly **higher** than the original national average. Those who took the tests in both years were therefore capable of making full progress in the intervening 12 months.

So what might have been happening in the wider context to account for the slow progress? A look back at the trend of performance in Year 3 between 1987 and 1995 may help to illuminate this. Between 1987 and 1991, the average reading performance of pupils in Year 3 fell slightly but significantly (Gorman and Fernandes, 1992); but then between 1991 and 1995, it rose again, and returned to the 1987 level (Brooks, Nastat *et al.*, 1996). Two significant aspects of the educational scene between 1987 and 1991 were that

- the National Curriculum was introduced, and
- the proportion of teachers in primary schools leaving their job during a year increased sharply, from 9 per cent in 1987 to a peak of 14 per cent in 1990, the second largest group of leavers being those taking early retirement (Dean, 1996).

The National Curriculum broadened the curriculum for early years, and in practice meant that **less** time was given to literacy. Simultaneously, the departure of an unusually large number of teachers from the profession may have led to a drop in the overall effectiveness of teaching, since much research confirms the commonsense belief that longer-serving and more experienced teachers are on the whole more effective, and that continuity is important for the quality of the teaching pupils receive. Both factors may have contributed to the fall observed between 1987 and 1991.

Between 1991 and 1995, both factors eased. The National Curriculum was revised, was no longer so crowded, and became more familiar to teachers, especially in Key Stage 1. Also, teacher turnover fell, and in 1993 was just under 8 per cent. This might be enough to account for the return of average reading performance in Year 3 to the 1987 level by 1995.

But in that case, why did this benefit not carry through into Year 4? Two continuing facets of the National Curriculum may be relevant, namely that

- National Curriculum assessments, in particular the tests for the end of Key Stage 1, occur in Year 2, and it may be that many schools concentrate their efforts on seeing that Year 2 pupils achieve the best National Curriculum test results that they are capable of. The effects of this might well continue to be apparent in Year 3;
- But on the other hand, in Key Stage 2 the National Curriculum took longer to become familiar; and its requirements become rather broad, with increasingly specific and time-consuming content to be covered across the (now) 10 statutory subjects (in England; 11 in Wales). This might mean that time for specific attention to reading is reduced; and inspection evidence in recent years has suggested a loss of pace and challenge in the earlier part of Key Stage 2 (OFSTED, 1995, 1996). If so, this might be enough to explain the slow progress between Year 3 and Year 4 in this study.

4.4 Implications

Because of the dearth of previous surveys at age 9, there is no way of knowing how the international result might have compared with other years. Previous downturns in national results in Britain have turned out to be blips in a generally stable graph, for example the drop in average reading scores in Year 3 between 1987 and 1991 already mentioned; and average reading attainment in Years 6 and 11 has changed remarkably little since monitoring began in 1948 (Brooks *et al.*, 1995). To get a more complete picture of trends over time, national monitoring surveys need to be carried out regularly; and when the opportunity arises for Britain to become involved in international surveys, those opportunities should be taken.

But the virtual absence of information on trends over time at age 9 should not lead to any downplaying of the national result on this occasion, namely the slow progress between Year 3 and Year 4. An informed and reasoned debate is therefore needed on the causes, and on remedies.

Remedies cannot be dispassionately agreed upon and implemented if the search for causes is equated with apportioning blame; no purpose would be served by laying all responsibility on, for example, teachers or teaching methods. Teachers have had to cope with significant organisational and curricular changes in recent years. Since all surveys of methods of teaching reading show that the great majority of teachers use a mixture of methods, it is extremely unlikely that pedagogical differences could influence overall results. And since average levels of performance have remained much the same since 1948, there is no evidence that 'traditional' methods succeeded and 'modern' methods fail.

Indeed, the stability of the long tail over time and across curriculum areas would seem to betoken a stubborn underlying tendency, namely that the British **educational system pays too little attention to low performers, and could and should pay them much more.**

It would seem to follow that one very effective way of raising our overall performance, in literacy and in other areas of the curriculum, would be to concentrate on those pupils who do not develop a reasonable functional level. To boost their achievement would be good both for them as citizens and for the country.

The culture of paying too little attention to low performers therefore needs to be changed.

What is needed to change it is

- the conviction among all partners in education that the vast majority of children can learn to read at least satisfactorily;
- for all partners in education to work together to remedy the problem – 'all partners' here meaning central government, Local Education Authorities, schools, governors, teachers, parents and children;
- a focused debate on whether the curriculum in the earlier part of Key Stage 2 is still too crowded and, if so, urgent moves to lighten the burden;
- fundamental research on 'what works' in raising literacy standards, especially for those at risk of failure;
- dissemination of effective strategies for keeping up the momentum in literacy in Years 3 and 4, and beyond;
- a broad mix of strategies, approaches and initiatives for improving literacy across the board;
- enabling initiatives on the part of national and local government; that is, support programmes designed to allow teachers to focus on literacy improvement, and with consistency in this support over time;

- early identification of children at risk of reading failure, where 'early' means by age 6 at the latest, followed by effective remediation;
- even earlier identification of children from families with a history of low literacy, so that those children can be given a preschool boost to prevent reading failure occurring.

Some of the effective strategies are already known, for example:

- *Cognitive Profiling System* (Kirkman, 1996)
- collaborative reading, as in the Avon Collaborative Reading Project (Gorman *et al.*, 1993)
- Education 2000 'Read It' Project (Sinson, 1995)
- Family Literacy, as promoted by the Basic Skills Agency (for evidence on its effectiveness, see Brooks, Gorman *et al.*, 1996)
- Phonological Training (Hatcher *et al.*, 1994; Sylva and Hurry, 1995)
- Reading Recovery (Sylva and Hurry, 1995)
- Some systematic phonics instruction in the early stages of learning to read (Adams, 1990)
- Success for All (Slavin *et al.*, 1994)
- Talking Computers in Education (Jersey Advisory Service, 1993)
- using trained reading volunteers effectively, as in the Knowsley Reading Project (Brooks, Cato *et al.*, 1996).

Different options exist and can be applied intelligently to different schools. In many cases, it is not a matter of applying one solution, but several **in combination**. Most generally of all, **it is time to start treating reading failure as preventable, and to commit the determination and the resources to prevent it.**

REFERENCES

ADAMS, M.J. (1990). *Beginning to Read: Thinking and Learning about Print.* Cambridge, MA: MIT Press.

ADULT LITERACY AND BASIC SKILLS UNIT (1995). *Older and Younger: the Basic Skills of Different Age Groups.* London: ALBSU.

BINKLEY, M. and RUST, K. (Eds) (1994). *Reading Literacy in the United States: Technical Report of the U.S. Component of the IEA Reading Literacy Study.* Washington, DC: US Department of Education.

BROOKS, G., FOXMAN, D. and GORMAN, T.P. (1995). *Standards in Literacy and Numeracy: 1948-1994* (NCE Briefing New Series No.7). London: National Commission on Education.

BROOKS, G., CATO, V., FERNANDES, C. and TREGENZA, A. (1996). *The Knowsley Reading Project: Using Trained Reading Helpers Effectively.* Slough: NFER.

BROOKS, G., GORMAN, T.P., HARMAN, J., HUTCHISON, D. and WILKIN, A. (1996). *Family Literacy Works: the NFER Evaluation of the Basic Skills Agency's Family Literacy Demonstration Programmes.* London: Basic Skills Agency.

BROOKS, G., NASTAT, P. and SCHAGEN, I. (1996). *Trends in Reading at Eight.* Slough: NFER.

DEAN, C. (1996). 'London sees rise in resignations,' *Times Educ. Suppl.*, **4168**, 17 May, 8-9.

EKINSMYTH, C. and BYNNER, J. (1994). *The Basic Skills of Young Adults.* London: Adult Literacy and Basic Skills Unit.

ELLEY, W.B. (1992). *How in the World do Students Read?* Hamburg: International Association for the Evaluation of Educational Achievement.

GORMAN, T.P. and FERNANDES, C. (1992). *Reading in Recession.* Slough: NFER.

GORMAN, T.P., HUTCHISON, D. and TRIMBLE, J. (1993). *Reading in Reform: the Avon Collaborative Reading Project.* Slough: NFER.

GORMAN, T.P., PURVES, A.C. and DEGENHART, R.E. (1988). *The IEA Study of Written Composition I: the International Writing Tasks and Scoring Scales.* Oxford: Pergamon Press.

HATCHER, P., HULME, C. and ELLIS, A.W. (1994). 'Ameliorating early reading failure by integrating the teaching of reading and phonological skills: the phonological linkage hypothesis,' *Child Development*, **65**, 1, 41-57.

JERSEY ADVISORY SERVICE (1993). *The Jersey Computer-assisted Reading Development Programme.* Jersey: Jersey Advisory Service.

KIRKMAN, S. (1996). 'Alien mission to find earthling dyslexics' (Primary Update). *Times Educ. Suppl.*, **4171**, 7 June, 10-11.

KISPAL, A., GORMAN, T.P. and WHETTON C. (1989). *Reading Ability Series Levels A-F, Teachers' Handbook.* Windsor: NFER-NELSON.

OFFICE FOR STANDARDS IN EDUCATION (1995). *English: a Review of Inspection Findings 1993/94.* London: HMSO.

OFFICE FOR STANDARDS IN EDUCATION (1996). *The Teaching of Reading in 45 Inner London Primary Schools. A Report by Her Majesty's Inspectors in Collaboration with the LEAs of Islington, Southwark and Tower Hamlets.* London: OFSTED.

SINSON, J. (1995) *Evaluation of Education 2000 'Read It' Project.* Leeds: Leeds Education 2000.

SLAVIN, R.E., KARWEIT, N.L. and WASIK, B.A. (1994). *Preventing Early School Failure.* Boston: Allyn and Bacon.

SYLVA, K. and HURRY, J. (1995). *Early Intervention in Children with Reading Difficulties: an Evaluation of Reading Recovery and a Phonological Training. Full Report.* London: SCAA.

THORNDIKE, R.L. (1973). *Reading Comprehension in Fifteen Countries.* Stockholm: Almqvist and Wiksell.

TIMES EDUCATIONAL SUPPLEMENT (1992). 'England decides it is beyond compare', *Times Educ. Suppl.*, **3981**, 16 October, 14.

APPENDIX A:
FULL DESCRIPTION OF HOW THE SURVEY WAS CARRIED OUT

A.1 The tests used

The tests used in this survey were:

- slightly adapted versions of the two parts of the test used with 9-year-old pupils in 27 countries in the 1991 IEA Reading Literacy study (Elley, 1992). The versions used in England and Wales in 1996 are described in Appendix C, and the adaptations are described and explained there. The vocabulary section of the IEA test contained 40 multiple-choice items, and the main sections, taken together, contained 68 comprehension items, of which 62 were multiple-choice and six were open-ended. The vocabulary section and eight of the 68 items in the main test were excluded from the main analysis in the 1991 study, and not reported there or in this survey;
- level B of the *Reading Ability Series*. A brief description of this British test is given in Appendix B, and full details are available in the *Teachers' Handbook* to the series (Kispal *et al.*, 1989). The total number of items is 35, of which 23 are multiple-choice and 12 are open-ended.

All the pupils in this survey took *RAS* level B, and the vocabulary section which was common to the adapted versions of the two parts of the IEA test. Then two equivalent half-samples of pupils took the two main parts of the IEA test. The two equivalent half-samples were achieved by allocating IEA test version 1 to pupils with odd numbers, and IEA test version 2 to those with even numbers; this was done to ensure that the half-samples would be equivalent, and their equivalence was later checked statistically (see section A.8, especially Table A.4, below).

In order to avoid order-of-presentation effects, half of the schools involved were asked to administer the IEA test first, and the other half to administer *RAS* level B first.

Administration times for the tests were as follows:

- the IEA test required one session, in which the two versions were administered simultaneously to the two half-samples. The sessions consisted of 1½ minutes for the vocabulary section, plus 40 minutes for the main section, plus a few minutes for setting up and collecting in. The IEA session took place at least a day before the first *RAS* session, or at least a day after the second one;

- *RAS* level B required two sessions, each consisting of 45 minutes' working time, plus a few minutes for setting up and collecting in. The two sessions were held at least a day apart.

The time taken up by the survey per pupil was therefore about 150 minutes.

A.2 The sample of schools

The survey was carried out in a nationally representative sample of schools in England and Wales containing Year 4 pupils. The school sample used had already participated in a survey of Year 3 pupils one year earlier (Brooks, Nastat *et al.*, 1996). On that occasion, there were four samples of schools, and of those the fourth and largest (which outnumbered the other three put together) was itself a nationally representative sample of 72 schools and 2174 pupils; this is the sample which was approached again in 1996.

In order to select the schools in that sample in 1995, the stratifying variables shown in Table A.1 were used.

Table A.1: Stratifying variables used to select schools

School type (independent/Infant/Junior/Infant and Junior)
Proportion of pupils receiving free school meals (within maintained schools)
Size of Year 3 age group
Type of LEA (metropolitan/non-metropolitan)
Region (North, Midlands and South of England; Wales)

All 72 of the schools which had taken part in 1995 were asked to take part in 1996; 58 schools agreed to do so and returned both the tests and full pupil information, giving a response rate of 81 per cent. The representativeness of the achieved sample was checked by comparing the proportions of schools in various categories in the sample with the national distribution; the results are shown in Table A.2.

Table A.2: Distribution of schools in various categories, in sample and nationally

	Population		Schools used	
	%	Number	%	Number
School type				
Independent	9	1587	7	4
First	10	1678	9	5
Junior	13	2332	17	10
Jun. & Infant	68	12057	67	39
Size of year group				
1-30	54	9518	52	30
31+	46	8136	48	28
Type of LEA				
Metropolitan	30	5268	31	18
Non-metropolitan	70	12386	69	40
Region				
North	30	5281	24	14
Midlands	23	4024	28	16
South	39	6911	40	23
Wales	8	1438	9	5

Since percentages are rounded to the nearest integer, they may not always sum to 100.

In all cases the proportions were very close. The distributions of schools in the population and in the sample with differing proportions of pupils receiving free school meals were also checked and found to be in close agreement. The achieved sample of schools was therefore considered to be adequately representative.

A.3 The sample of pupils

The pupils involved in 1996 were all in Year 4, and were born between 1 September 1986 and 31 August 1987. At the date of testing they were aged between 8 years 6 months and 9 years 6 months old, and the average age was 9 years 0 months (9:00). Since they were on average 8.4 months younger than the pupils who had taken part in the 1991 IEA study, an age correction was applied to the results of the IEA test – see chapter 3.

At an early stage in this study, a decision had to be made on whether to test in Year 4 (age 9) or Year 5 (age 10). The definition of the target population in the IEA report (Elley, 1992, p.101) was

> *All students attending mainstream schools on a full-time basis at the grade level in which most students were aged 9:00-9:11 years during the first week of the eighth month of the school year.*

The 'eighth month of the school year' for England and Wales is April. Since school cohorts here are defined by date of birth, with cohorts beginning on 1 September in one year and ending on 31 August in the next, on 1 April in any calendar year

- Year 4 pupils are aged between 8:07 (8 years 7 months) and 9:06, with an average of 9:00, while
- Year 5 pupils are aged between 9:07 and 10:06, with an average of 10:00.

Thus the international definition of the target population comes down, for England and Wales, precisely at the dividing line between Years 4 and 5.

To produce a sample whose age range was 9:00-9:11 and whose average age was exactly 9:06, schools could have been asked to test, in March 1996, pupils with dates of birth between 1 March 1987 and 29 February 1988, of whom approximately half would have been in Year 4 and half in Year 5. But in addition to breaching the IEA's guideline of testing **within** one school grade or year, this would have been much more awkward for the schools. The choice was therefore between Years 4 and 5 as a whole. And then to have tested Year 5 would have meant passing up the chance to include a longitudinal aspect by testing the pupils who had already been tested in Year 3 in 1995. By extension, it would also have been impossible to make comparisons with the two previous surveys at age 8 (1987, 1991). Hence the decision to test in Year 4.

In each participating school, all the pupils in Year 4 took part in the survey. The number of pupils per school ranged between a handful and over 100, with an average of 31.

The purpose of returning to the schools which had taken part in 1995 a year later was to re-test as many as possible of the same pupils, and this objective was achieved. The numbers of pupils for whom tests were returned, including the number who participated in both years, are shown in Table A.3.

Table A.3: Numbers of pupils for whom tests were returned

Test	Number of pupils
IEA test	
– part 1	900
– part 2	917
– Total	1817
***RAS* level B**	
– Total	1803
– Pupils who had also taken level A in 1995	1504

It should be noted that the numbers of pupils for whom results are reported in chapter 3 are not always the same as those just given; this is due to missing information, for instance on whether or not pupils were receiving free school meals.

A.4 The survey design

The survey was designed to permit several forms of comparison:

1. between performance in England and Wales in 1996 and 27 other countries in 1991, on the IEA test;

2. between the IEA test and *RAS* level B;

3. between performance on *RAS* level B in 1996 and its standardisation in 1987;

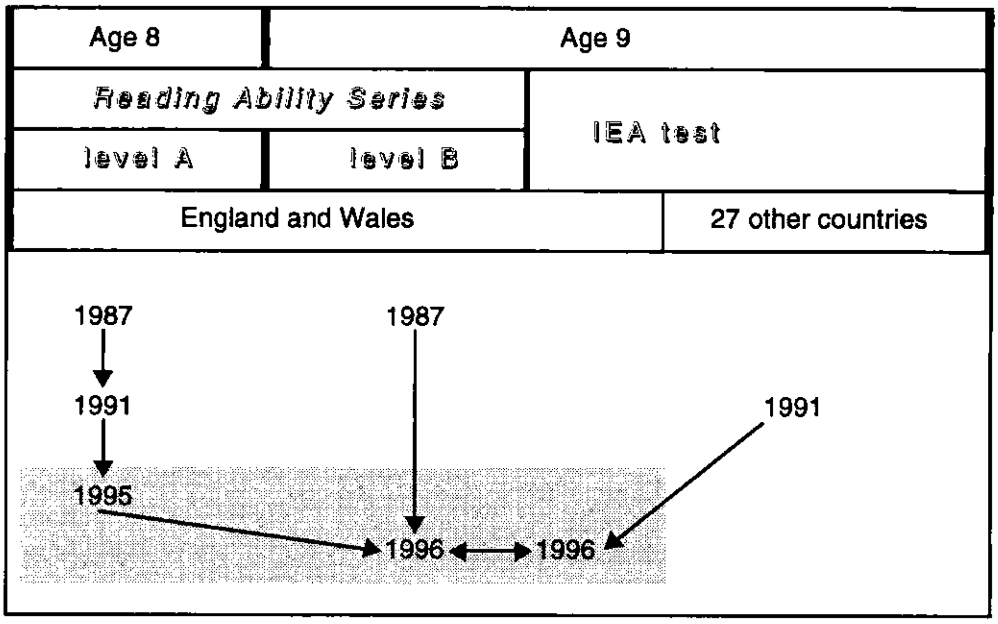
4. between performance on *RAS* level B in 1996 and level A in 1995. This was possible because the great majority of pupils involved in 1996 had also participated in 1995, and this was therefore a small-scale longitudinal aspect of the study.

Comparisons 1 to 3 just listed were based on the whole sample of pupils tested in 1996; the longitudinal comparison between 1995 and 1996 (no. 4), however, was of course based only on those who took part in both years.

It was also intended to relate the data gathered in 1996 to information on the trend of reading standards at age 8 (Year 3) between 1987 and 1995. This was possible via the link between level B in 1996 and level A in 1995, and because surveys at age 8 in 1987 and 1991 had also used *RAS* level A (Gorman and Fernandes, 1992; Brooks, Nastat *et al.*, 1996).

The set of comparisons is schematised in Figure A.1.

Figure A.1: Schema of the survey comparisons



Key: Boxing indicates same pupils involved

A.5 Date of testing

The survey took place in March 1996. This was, by design, exactly one year after the Year 3 survey, and the time of year at which most of the 1991 IEA testing was done. In this way, time-of-year effects were avoided.

A.6 Background data

Most of the background information needed for this survey was already available, having been gathered in 1995. In 1996, headteachers were asked only to

- state, for all pupils, whether or not they were receiving free school meals
- state date of birth and sex of pupils who had not taken part in 1995.

It should be noted, therefore, that the study was designed simply to investigate where England and Wales might have come in the 1991 IEA study, and to compare that result with the outcome of a British test. It was **not** designed to shed light on relationships between the test results and background factors (other than sex and receipt of free school meals); in particular, the effects on pupils' attainment of different approaches to the teaching of reading were not investigated - that would have required a very different study.

A.7 Coding and marking the tests

Within the IEA tests, only the six open-ended questions required coding. Four of these questions required short written answers and two a longer response. The two requiring a longer response were left for coding and analysis at a later stage. The four short-answer questions were coded, and are included in those on which this report is based.

The *RAS* tests were marked by a team of four experienced markers known to NFER using the printed marking key. The markers' reliability was checked by Anne Kispal, one of the authors of the test, and found to be satisfactory.

A.8 Checking the equivalence of the half-samples for the IEA test

The most significant difference in procedure on the IEA test between the 1991 study and the 1996 survey was that

- in the 1991 IEA study, all pupils took both main parts of the IEA test as well as the vocabulary test;
- in England and Wales in 1996, the two main parts of the IEA test were taken by separate half-samples of pupils, and only the vocabulary test was taken by all of them.

The decision to make this change in procedure was taken to avoid subjecting each pupil to a fourth test session. Because of the difference in procedure, before comparing the 1996 England and Wales results with those of the 1991 study it was necessary to establish that the two half-samples of pupils were equivalent, and therefore that their results could be summed. This was done by calculating the average raw scores and standard deviations for the two half-samples, and various correlations and reliability coefficients. The results are shown in Table A.4.

Table A.4: Average raw scores, standard deviations, correlations and reliability coefficients within the IEA test

Average raw score (and s.d.) for vocabulary test for half-sample		
	A:	33.0 (8.8)
	B:	32.3 (8.8)
Correlation between vocabulary test and main test		
	part 1:	0.92
	part 2:	0.85
KR20 reliability coefficient for		
– vocabulary test		0.97
– main test part 1		0.90
– main test part 2		0.93
– vocabulary test plus main test part 1		0.96
– vocabulary test plus main test part 2		0.96

Since the average raw scores and standard deviations were very close, and all the correlations and reliability coefficients were very high, it was concluded that the two half-samples were equivalent, and that their results could legitimately be summed to provide overall results for the IEA test.

A.9 How the test results were analysed

For the *RAS* test, each pupil's raw score (number of items correct) was converted into a standardised score using the tables in the *Teachers' Handbook* (Kispal *et al.*, 1989). These tables make allowances for differences in the ages of the pupils tested.

For the IEA test, the methodology followed was exactly that used by the IEA in 1991. First, a raw facility value was calculated for each item. In calculating the results in the 1991 IEA study, six items (all multiple-choice; two from each of the three Domains) which did not fit the statistical model used (see below) were dropped, and these were also excluded from the 1996 England and Wales calculations, leaving 60 items.

Next, the raw facility values for these 60 items were summed, to give an average raw score for the test as a whole. Average raw scores were also calculated for the two parts of the test.

Third, the data were processed using

- a one-parameter Item Response Theory (Rasch) model;
- the reporting scale used by IEA, which has an international mean of 500 and an international standard deviation of 100; and
- the same item parameters as in the 1991 study.

This process was used to calculate, in turn:

- Rasch scores for individual pupils, for each of the three types of text, or 'Domains', contained in the test (Narrative, Expository and Documents)
- average Rasch scores for each of the three Domains
- the average Rasch score for the whole test - this was calculated by taking the arithmetic mean of the average scores for the three Domains, and was, in effect, the overall score for England and Wales.

The Rasch statistical model assumes that certain postulated traits underlying human behaviour are unidimensional, and is controversial. The main reason for this appears to be that the model provides no mechanism for testing its own assumptions, particularly unidimensionality. One indication of this is that when test items do not fit the assumption of unidimensionality the model cannot deal with them, and they have to be excluded. Indeed, as already stated, in the 1991 study six items were excluded on these grounds. The Rasch model was nevertheless used with the IEA test data in this study (including the dropping of the six non-fitting items) because it had been used in the 1991 study, and direct comparability could be achieved only in this way.

A.10 A possible ceiling effect

It was explained above that no pupil took the whole of the IEA test; instead, the two parts were taken by equivalent half-samples of pupils. However, this aspect of the testing procedure for this study had an unintended consequence, which arose from an interaction of this aspect of the procedure with the statistical method used to calculate overall results. In the IEA study, each pupil's Rasch score for each of the Domains was based on about 20 items (see Table 2.1). In this survey, each pupil's score for a particular Domain was based on only the items from that Domain which happened to be in the part of the test which the pupil had taken. The numbers of items on which pupils' Rasch scores for the Domains were based were therefore lower (see again Table 2.1); in particular, the Documents Domain in part 1 contained only six items. The distribution of scores on the Domains represented by few items showed a clear ceiling effect; in some cases, the most frequent score was the maximum score. It follows that some pupils did not have sufficient scope to achieve scores well above the average of their peers. This will have depressed both the averages for the Domains and the overall average score to an extent which is unquantifiable. The implications are discussed in the main body of this report.

A.11 Feedback to schools

It was part of the project design that, once marking was complete, each of the schools involved would be sent the test results of those of its pupils who had participated. This information is to be sent out soon after the publication of this report.

APPENDIX B:
Reading Ability Series

This consists of six levels, A-F, which between them provide standardised norms for children of ages 7:00 to 13:11. There is an associated *Test of Initial Literacy* suitable for weaker readers at all these ages and for children under seven. The whole series was standardised in England and Wales in 1987; the questions are a mixture of multiple-choice and open-ended.

Each level of the series consists of a Reading Book and a Work Book. The Reading Book contains both a narrative text and an expository text. In levels B to F, the narrative comes first. The Work Book contains the questions on both texts, and the narrative and expository tests are taken in separate sessions, of 45 minutes each.

In level B, the narrative is called 'Uncle Charlie's Ramshackle Car' and has 17 questions (13 multiple-choice, four open-ended); the expository text is about Elephants and has 18 questions (10 multiple-choice, eight open-ended); there are therefore 35 questions in all (23 multiple-choice, 12 open-ended).

Like all the levels in the Series, level B is commercially produced in two tones, with a multi-coloured glossy cover to the Reading Book. The printing is very clear, and the test looks very up to date.

APPENDIX C:
THE 1991 IEA TEST

The test used in the IEA survey of 9-year-olds in 1991 had two parts, and all the pupils involved took both parts, in separate sessions.

The first part consisted of

- two practice items for the vocabulary section
- 40 multiple-choice picture vocabulary items
- two practice texts and six practice items for the main, comprehension, section
- six texts and 26 items constituting the comprehension section.

The second part consisted of nine texts and 42 test items. There were no practice items in the second part because all those taking it were assumed to have taken the first part.

Administration times were:

- for the first part, 1½ minutes for the vocabulary section, and 35 minutes for the main section
- 40 minutes for the second part.

There were also separate questionnaires to be completed by pupils, teachers and schools.

When the 1996 NFER/Open University study was being planned, several modifications to the IEA design were decided upon.

First, since almost all the background information which it would be feasible to use was already available from the 1995 Year 3 survey, it was decided to dispense entirely with the teacher and pupil questionnaires, and to use a very simple school questionnaire to gather only the information on individual pupils which needed to be up to date (free-school-meals status for all, sex and date of birth for those new to the schools).

Secondly, since all the pupils involved were to be asked to take *RAS* level B, which requires two sessions of about 50 minutes, it was considered unreasonable to require all the pupils to take both parts of the IEA test (two further sessions of about 45 minutes each) in addition. It was therefore decided that

- each pupil would take only one of the two main parts of the IEA test;
- but all pupils would take the vocabulary section. This would be used to check the equivalence of the two pupil half-samples, which would otherwise have relied solely on the *RAS* test;
- there would be just one practice text, with two items, between the vocabulary and main sections;
- test version 1 would therefore consist of the vocabulary section, first practice text and main section from part 1 of the IEA test (the only deletion being the second practice text);
- test version 2 would consist of the vocabulary section and first practice text copied over from part 1, followed by the (unaltered) main section of part 2 of the IEA test;
- since both versions were to be administered simultaneously in the same classrooms, to avoid having to stop pupils who would be taking version 1 five minutes earlier than those taking version 2, the time for the main section of both tests was set at 40 minutes (no change for version 2, an increase of five minutes for version 1).

The IEA test had been specially produced for the 1991 study, and not to commercial standards. The entire contents (text and illustrations) were in black and white. The type faces used seemed slightly old-fashioned. Moreover, the copies supplied as masters for this survey (which appeared to be of the copy standard used in 1991) were not quite as crisp, in print quality, as a commercially produced test would have been.

nfer

The Open University

READING PERFORMANCE AT NINE

In March 1996, the National Foundation for Educational Research (NFER) and the Open University jointly carried out a survey of reading attainment in Year 4 classes (pupils aged 9) in England and Wales. Most of the pupils had also been tested in Year 3 in 1995.

The survey was principally funded by the Esmée Fairbairn Charitable Trust, with additional support from Channel 4 Television, the *Mail on Sunday* newspaper, the NFER and the Open University.

Two tests were used:

- the test used in 1991 in a survey of 9-year-olds in 27 other countries;
- level B of the British *Reading Ability Series*.

The key findings were that

- the average score on the international test would have put England and Wales close to the overall average in the 1991 study, within a group of 13 countries whose average scores did not differ significantly;
- the average score for England and Wales on the international test was lowered by (among other factors) a 'long tail' of pupils who achieved scores well below the average; and
- the pupils tested in both 1995 and 1996 appeared to have made slower progress, on average, in the intervening 12 months than children did in 1987.

These findings are important for all partners in British education, from central government to pupils. They are presented here as material for an informed and reasoned debate about preventing reading failure.

ISBN 0 7005 1431 7
£6.50