The Primary Review

... children, their world, their education

INTERIM REPORTS

Research Survey 4/2

STANDARDS IN ENGLISH PRIMARY EDUCATION: THE INTERNATIONAL EVIDENCE

Chris Whetton, Graham Ruddock and Liz Twist
National Foundation for Educational Research

For other interim reports in this series, and for briefings on each report, go to www.primaryreview.org.uk

This report has been commissioned as evidence to the Primary Review. The analysis and opinions it contains are the authors' own.

Esmée Fairbairn FOUNDATION

UNIVERSITY OF CAMBRIDGE
Faculty of Education

The Primary Review

... children, their world, their education

PRIMARY REVIEW INTERIM REPORTS

STANDARDS IN ENGLISH PRIMARY EDUCATION:
the international evidence

Primary Review Research Survey 4/2

Chris Whetton, Graham Ruddock and Liz Twist

October 2007

This is one of a series of 32 interim reports from the Primary Review, an independent enquiry into the condition and future of primary education in England. The Review was launched in October 2006 and will publish its final report in late 2008.

The Primary Review, supported by Esmée Fairbairn Foundation, is based at the University of Cambridge Faculty of Education and directed by Robin Alexander.

A briefing which summarises key issues from this report has also been published. The report and briefing are available electronically at the Primary Review website: www.primaryreview.org.uk. The website also contains Information about other reports in this series and about the Primary Review as a whole. (Note that minor amendments may be made to the electronic version of reports after the hard copies have been printed).

We want this report to contribute to the debate about English primary education, so we would welcome readers' comments on anything it contains. Please write to: evidence@primaryreview.org.uk.

The report forms part of the Review's research survey strand, which consists of thirty specially-commissioned surveys of published research and other evidence relating to the Review's ten themes. The themes and reports are listed in Appendices 1 and 3.

This survey relates to Primary Review theme 4, **Quality and Standards**.

The authors are at the National Foundation for Educational Research (NFER). Chris Whetton is Assistant Director; Dr Graham Ruddock is a Deputy Head of the Department for Research in Assessment and Measurement; Dr Liz Twist is a Principal Research Officer.

Suggested citation: Whetton, C., Ruddock, G. and Twist, L. (2007) *Standards in English Primary Education: the international evidence* (Primary Review Research Survey 4/2), Cambridge: University of Cambridge Faculty of Education.

Published October 2007 by The Primary Review,
University of Cambridge Faculty of Education,
184 Hills Road, Cambridge, CB2 8PQ, UK.

British Library Cataloguing in Publication Data:
A catalogue record for this publication is available from the British Library.

ISBN 978-1-906478-02-5

STANDARDS IN ENGLISH PRIMARY EDUCATION: THE INTERNATIONAL EVIDENCE

Introduction

This review examines international survey evidence on the performance of English children of primary school age in relation to those from other countries. It starts by setting out the context of these international surveys and specifies those which are discussed here. It examines the methodological basis of the surveys, noting criticisms and problems, before considering the survey findings in mathematics, reading and science. The strengths and limitations of the data are assessed, and implications for the future international monitoring of educational standards are identified.

The context

International comparative studies of educational achievement began in the early 1960s, in part as a cold-war reaction to the Soviet Union's launch of the first orbital satellite and the consequent concerns about levels of technical skills. The questioning of education systems which resulted lay behind the first international mathematics study in 1964. Early surveys were long-drawn out studies held at irregular intervals and with methodological weaknesses. (An early history is given in Husén and Tuijnman, 1994.) In contrast, modern surveys are tightly conducted, relatively rapid in reporting, involve more countries and are at regular intervals, allowing time sequences of information. They are also robust in methodological terms, though still not without critics of their operation and underlying philosophy (see, for example, Bonnet 2002 or Hilton 2006). Like the early studies, the current surveys operate in a political context but this is now that of global competition and a believed link to economic prosperity (Bonnet 2002). An overview of the purposes and conduct of surveys is given by Beaton *et al.* (1999).

There are currently two main sets of international surveys: those conducted by the International Association for the Evaluation of Educational Achievement (IEA) and those conducted by the Organisation for Economic Cooperation and Development (OECD) as a part of its activities, known as the Programme for International Student Assessment (PISA). The IEA is an international non-governmental organisation whose members are research centres and ministries of education. Its studies are designed to inform researchers, educators, policy makers and the public about educational achievement, and to relate this to contextual factors. The IEA studies have their roots in educational research and, since the founders of the organisation tended to be academic research centres, the approach tends to be bottom-up, defining the context of its tests through the communality of participating countries' subjects and curricula. The studies discussed here are currently conducted by an International Study Centre, based in Boston College, USA. A general description can be found at www.iea.nl.

PISA is overtly steered by the governments of the members of OECD, although other countries can participate. As such it is more reactive to the desires of national policy makers and this is reflected in its approach. This has been to define the skills needed for the populations of modern economically advanced countries and then to assess the extent to which these are present, independently of the countries' curricula. For this reason, PISA tests students at or near the end of schooling, and does so in three 'literacy' areas of reading,

mathematics and science. The programme undertakes surveys every three years, rotating the subjects so that every nine years one particular area is the main focus and the others subsidiary. In 2000, the main focus was reading literacy; in 2003 it was numeracy; and in 2006 (yet to report) it was science. The sequence will then repeat from 2009 onwards. A general description can be found at www.pisa.oecd.org.

For a short time in two years (1989 and 1992) (Lapointe 1992a and b), there were also two large-scale international studies undertaken by the US's Educational Testing Service, under the name 'International Assessment of Educational Progress' (IAEP). These covered mathematics and science, but only the second included primary schools. For this review, we have excluded other smaller international comparative studies. This is because they generally involve a single comparison with one other country and cannot provide a wide context, or because they were based on opportunity samples which would not meet the criteria of the full surveys, which utilise careful checks on samples and their attainment. (For examples see Martin *et al.* 2003, 2004a.)

Neither PISA, with its focus on school leavers, nor the IEA studies have concentrated on primary schooling, and hence the information for this age group from international comparative studies is relatively sparse.

To date, there have only been six reputable international studies of primary-aged children in which England has participated. These were:

IEA	Reading	1971
IEA	Science	1984
IAEP	Mathematics and science	1991
IEA	Mathematics and science (TIMSS)	1995
IEA	Reading literacy (PIRLS)	2001
IEA	Mathematics and science (TIMSS)	2003

Information is therefore rather sporadic and drawing strong conclusions is not advisable. A recent trend has been for the surveys to be held at regular intervals allowing more consistent data and a better understanding of changes taking place over time. It is likely, then, that better data will be available in the future. There has been a further PIRLS study in 2006, as yet unreported, and a TIMSS study in 2007.

For an earlier survey of England's (or the UK's) standards in literacy and numeracy to June 1994 see Brooks, Foxman and Gorman (1995). Reynolds and Farrell (1996) provide a summary of four major international comparisons covering 1960-91. A more recent review is Smithers (2004) for the Sutton Trust. None of these concentrated on primary education.

It should be stressed that measures of achievement are only part of these large international surveys. They also collect a great deal of contextual data. Included, for example, are information on children's attitudes, their home backgrounds, teachers' experience and qualifications; the nature of the school; the national educational system; and, in recent surveys, parents' views. As such the surveys offer huge opportunities for secondary analyses, and it is unfortunate that this expensively generated data has not been utilised to a greater extent. In part this may be because of the size and complexity of the data, but great efforts are now made to make it available for research scrutiny.

In examining achievement data, there are several issues to be considered. Of course most interest is immediately focused on the mean attainment measures, the absolute level, standing compared to other countries and changes over time. However, other aspects are also important beyond these headline measures. The spread of attainment is of interest, for example. Is this narrow, indicating a cohesive education system in which all attain around the same level (whether high or low), or is the spread large, indicating a wide range of attainment and great disparities between the highest and lowest attainers? Related to this, the levels achieved by the highest attainers (these should be high for a high-value research and development led economy) and the levels achieved by the lowest attainers (the baseline levels of literacy and numeracy in a country) are both of interest.

Methodological limitations

If conducting a survey on evidence of standards in one country is difficult, conducting one across many countries borders on the impossible. There are many sources of this difficulty which stem from the different underlying philosophies of education, the different structures of educational systems, the different curriculum emphases and, finally, the potential different languages. International surveys adopt a series of techniques to attempt to make these as comparable as possible.

In essence, the approach taken by the various surveys is similar, though the language and precise processes may differ. A first stage agrees the content framework for the surveys and the approach to assessment to be taken: the modes of testing and style of items. These should be widely discussed by participants and agreed through processes involving their representatives. The tests themselves are usually the responsibility of a single agency, but the best practice is to draw on contributions from a wide range of participating countries, originated in many languages. The draft items are formed into several alternative forms and administered in field tests, usually the year before the actual survey. These field tests have the purpose of trialling the items to ensure that they function well psychometrically in each country and do not perform very differently in any of the countries. The field tests also allow a rehearsal of the processes to be used in the subsequent main survey. For both the field test and the main surveys, participating countries translate the tests into their own language(s) of education and submit these to a translation verification process, which involves independent scrutiny of the translation and the level of language adopted. Translations into the same language from different countries are compared and aligned. The samples of schools and children for the surveys are either drawn by an independent agency or have to be verified by a sampling referee. The numbers utilised are substantial, generally running to thousands of children in hundreds of schools. There is frequently random selection of pupils within schools rather than complete cohorts. Stringent criteria for inclusion are set for both school participation and the percentage of selected children to be achieved. Countries not meeting these are excluded or distinguished from the remainder in some way. Scrutineers from within or outside the countries observe the testing on an unannounced basis to ensure it is being conducted as required. (All of these processes are documented in comprehensive manuals.) Following the administration of the tests, they are marked within countries using common scoring systems, but with the operation supervised by people who have been centrally trained in consistency at international meetings. Proportions of the tests are double marked to check reliability and there may also be verification processes where some tests from one country are remarked in another country. The data capture is generally done using the same software in each country with the same embedded verification processes. There are generally many versions of the tests but arranged in systematic patterns so that they have some common questions, allowing the whole survey to be scaled together, the items to be calibrated on to one scale and the pupils to have their attainment measured in a comparable

manner. This is done by a central analysis agency, which also examines the data for biases in any questions in particular countries or for questions which have performed differentially for reasons of the translation or otherwise. Finally the data are compiled into international and national reports, which include statements about the significance of the data, the differences between countries and the relationships to contextual variables. The processes of each survey are thoroughly documented in publications open to scrutiny.

Undertaking these operations is a detailed and onerous task, so international comparative studies of educational standards are large and expensive exercises. As such, it is reasonable to question the validity and reliability of the results they produce. There have been many critical examinations of these studies which to a greater or lesser extent suggest flaws in the methods adapted (Bonnet 2002, and Goldstein 2004 – of PISA; Clark 2004, and Hilton 2006 – of PIRLS; Galton 1998, and Winter 1998 – of TIMSS). Counter arguments have been put by Beaton *et al.* (1999), Whetton *et al.* (2007, forthcoming) and the studies themselves.

The criticisms can be grouped into four types: those that relate to the underlying conceptualisation of the studies as research enterprises; those that concentrate on cultural and linguistic factors; those that question the statistical and psychometric basis; and finally those which examine the sampling methodology.

The first set of criticisms relates to the conceptualisation of the studies. It is certainly the case that the recent motivation of many governments for participation in comparative studies is because of an assumed link between educational standards and economic success. Bonnet (2002) in particular has been critical of this assumption, arguing that the pursuit of causation is a chimera. Bonnet also suggested that the studies by their very nature accept a dominant model of schooling and enforce it on all even when the model does not apply, leading to incorrect conclusions. This viewpoint has been expressed most forcibly among French language commentators; see Lafontaine (2004) for example. The title of one paper sums up this view nicely: *Le bon (critique), la brute (médiatique) et les truands (anglo-saxons)* [The critical good, the rough media and the Anglo-Saxon gangsters.] (Lafontaine and Demeuse 2002). There is probably no argument that can be used to overcome these objections, except to say that they are a counsel of despair and, if taken to their conclusion, mean that no cross-cultural educational comparisons are possible; a view which would not be accepted generally. In general those conducting the surveys are aware of the issues and strive to overcome them.

In an attempt to address such concerns, a study funded by the European Union Socrates programme (Bonnet *et al.* 2001, 2003) explored an alternative methodology for international comparative surveys. This study looked into 'the feasibility of implementing an internationally comparable survey of pupils' attainment in reading based on the use of indigenous untranslated test instruments in order to lessen linguistic and cultural biases' (Bonnet *et al*. 2001). The impetus for this work was doubt about the possibility of devising assessment instruments without cultural bias, in addition to a view that the English language was unduly dominant in original materials (from which translations were made) in previous studies. The study involved educationalists from England, Finland, France and Italy. The methodology adopted attempted comparative analyses whilst using assessment materials in their original language. The basis for the study was the construction of the national instruments according to a common framework of skills, levels of difficulty, text types and item types. It required the use of a common anchoring test, which was calibrated in each participating country. This was the vocabulary sub-test of the Wechsler Intelligence Scale for Children version 3 (WISC III). Those involved concluded that this approach offered some promise but that considerable further work was needed, including greater detail in test specifications in relation to sampling, item construction, and more sophisticated data

analysis methods. However, they tended to overlook the contradictory fact that their methodology ultimately rested on intelligence tests, originating in an American context and subsequently translated, and therefore open to the same criticisms made of the standard survey methodology.

The next set of criticisms applies particularly to the tests of reading in PISA and in PIRLS, but also has some resonance in the testing of science. This argument is that linguistic and cultural factors make it impossible to compare countries fairly. The pre-existing knowledge of students is said to be such that they bring different assumptions to the situation. Again see Bonnet (2002) for this argument and Hilton (2006) for an English expression of it. Whetton *et al.* (2007, forthcoming) give a refutation in the context of PIRLS.

Bechger *et al*. (1998) go so far as to suggest 'validity within nations and comparability across nations may be conflicting aims' (p. 101). In fact arguments are made for the cultural-specificity of texts, not only between but also within countries. Whilst acknowledging that the development process itself, including piloting of the PIRLS tests undertaken during this phase, made the tests 'as culturally fair as possible', Hilton suggests that the underlying methodology 'ignores deep cultural differences both between nations and between different groups in each nation' and that attempts to reduce this cultural specificity results in poorer assessment tools. The argument is raised both in relation to the texts themselves and also the items. Hilton argues that the texts are 'drained of cultural specificity through trialling and elimination, they are in fact also leached of intrinsic interest, comprehensibility, and vitality' (2006, p. 824). Whetton *et al.* (2007, forthcoming) provide a detailed refutation of these views.

Related to the cultural criticisms is the issue of translation. The demands of translation are substantial in these international comparisons and are discussed by Bechger *et al.* (1998), Bonnet *et al.* (2001) and Blum *et al.* (2001) in relation to literacy assessments. In the 2001 PIRLS cycle, for example, the tests were translated from English into 31 other languages. The translation and verification of the resulting translated texts in international comparisons is an extremely thorough and well-documented process, see for example Kelly and Malak (2003), and in general works well in modern studies. However, there can be problems to which the developers need to be alert. Investigations in relation to the IALS survey into adult literacy in the mid-1990s raised a number of concerns about equivalence, and Blum *et al*. (2001) illustrate these in relation to specific items.

All the international surveys utilise a statistical method generally known as item response theory (IRT) (Van der Linden *et al.* 1997 for an overview). This technique scales the difficulty of questions in the tests and produces estimates of the ability of students. It is fundamental to the design of the studies, since it allows students to take different tests and their results to be combined through common or linking items. IRT is in general use throughout the world for psychometric studies but its use in international comparisons has been questioned, particularly by British critics (Goldstein 2004; Hilton 2006).

Goldstein criticises the international surveys for the lack of any systematic procedure for evaluating the IRT technique, and suggests that their data is in fact more complex than allowed for by IRT. He is particularly dubious about the assumptions of unidimensionality in IRT and the practice of removing items that do not fit the IRT models well. For reading, for example, this practice may serve to impose a pre-determined unidimensional model of reading achievement (Blum *et al.* 2001; Goldstein 2004; Hilton 2006). However, this may not actually occur. It is clear from the Technical Report for PIRLS 2001 (Mullis *et al.* 2003) that the items included within the final tests did support a unidimensional model in that just two items were identified as problematic, one because an incorrect mark scheme was applied in one country and one because of a translation error in one of the languages in one country.

These particular items were removed from the analysis for these countries only. No other items displaying large item-by-country negative interactions were identified (i.e. when a country's performance on an item was unexpectedly low, given its overall performance and the performance of other countries on that particular item). Similar results have been found in TIMSS surveys (eg Martin *et al.* 2004b).

The final criticisms relate to the specification and achievement of the samples of students, with suggestions that these may not be representative of the individual countries. Winter (1998), for example, argued that international studies do not take sufficient account of sampling problems when comparing different countries.

The issue of sampling is of critical importance in international comparisons. It is essential that the sampling method adopted provides an accurate sample from which the data can be derived, whilst remaining manageable across all participating countries and education systems.

It is important to note that in the modern studies there are strict sampling targets that individual countries must achieve in order to be included in the main tables of the international report, and that the sampling framework adopted has to be approved by an independent organisation. In the case of IEA, this has been Statistics Canada. For PISA it has been WestStat. Both of these are substantial institutions with a great depth of expertise. The consequence of not meeting one or more of these targets was shown by the exclusion of the United Kingdom from the PISA 2003 reports (OECD 2004).

As an example, the sample design implemented in the PIRLS 2001 assessment is generally referred to as a three-stage stratified cluster sample. *The first-stage sampling units* consist of individual schools. Schools are selected with probabilities proportional to their size (PPS); size being the estimated number of pupils enrolled in the target grade, year 5 in PIRLS in England. The comprehensive national list of all eligible schools is called the school sampling frame. As the schools are sampled, replacement schools are simultaneously identified should they be needed to replace sampled schools which decline to participate. *The second-stage sampling units* are classrooms within sampled schools. Within each sampled school, a list of eligible classrooms from the target grade is prepared. A single eligible classroom per target grade is randomly selected from each participating school. *The third-stage sampling units* are pupils within sampled classrooms. Generally, all pupils in a sampled classroom will be selected for the assessment.

There are various participation targets which must be met, not all of which were fully met by England in PIRLS 2001: 85 per cent of initially sampled schools, 95 per cent of sampled classrooms and 85 per cent of sampled students and teachers; or a minimum combined school, classroom and student participation rate of 75 per cent, based on sampled and replacement schools (Joncas 2003).

To be included in the international report with annotation, as England was in 2001, the sample must meet the above targets with the inclusion of replacement schools and include at least 50 per cent of initially sampled schools and have a school participation rate of at least 50 per cent.

It is at the first and third stages of the sampling that concerns have been expressed about the representativeness of the achieved sample for PIRLS 2001 (Clark 2004; Hilton 2006). England's weighted participation rate of sampled schools (i.e. 'first choice' schools) was 57 per cent; with replacement schools this increased to 88 per cent. At the third stage, the weighted pupil participation rate was 94 per cent. These participation rates led to England's inclusion in the international report with an annotation to indicate that replacement schools

were required to meet sampling targets and that the proportion of pupils participating was less than 95 per cent of the national desired population.

Ultimately, each interested person must make their own view on the reliability and validity of these international surveys and of the methodological criticisms made of them. But this cannot be a single view of them all. The early studies did have weaknesses: in test specification (eg IEA reading 1991); in sampling; in sample verification; and in analysis. These, though, have been learned from and addressed as far as possible in later surveys. The underlying constructs to be assessed are now published in framework documents (eg OECD 2006 for PISA; Mullis *et al.* 2006 for PIRLS; Mullis *et al.* 2005 for TIMSS). The test developers attempt to draw on material from many participating countries and to cover a range of cultural approaches. Although the studies continue to work in English, the translations of tests are checked and verified carefully and sensitively. The samples are drawn by independent sampling organisations, not the countries themselves, and their achievement is monitored and checked. The final sample ratios required are high, and countries failing to meet them are excluded from the published results. Independent monitors view a selection of the test administrations in every country. The analysis techniques are agreed by technical committees and implemented with checks for dimensionality and the functioning of items. The reporting is careful to state the significance (or lack of significance) of differences. All of this is a considerable and expensive validation process, but whether it is sufficient has to be a personal view. For some it can never be. Bonnet (2002) considers the cultural model to be flawed. Goldstein (2004) considers the statistical model to be flawed. The authors of this review are involved in various ways in international surveys and need to declare that interest. It is our view that the methodology of the surveys presented here is sufficiently robust that their results can be considered to give a reasonable impression of the performance of the students in a participating country, compared to those in the other countries.

Mathematics

The IEA First International Mathematics Study, in 1964, was the first important international comparative study of this subject area. It did not, however, involve primary-aged pupils and this was also the case for the Second International Mathematics Study, in 1980-82. A different organisation, the International Assessment of Educational Progress, then mounted an international study of mathematics performance in 1988, but again this concentrated only on the secondary age range. A second study from this organization did involve the primary age group and this study, in 1991, provides the first systematic information on how primary mathematics performance in England compared with that in other countries.

Mathematics: The 1991 IAEP study

The parent organisation of the 1991 IAEP study was Educational Testing Service (ETS) of the USA (Lapointe *et al.* 1992b). The target age group was 9 year olds, and pupils from England were drawn from the two year groups containing pupils of this age.

The response rate for schools in England was 56 per cent, the lowest of all the participating countries, but not much below Scotland. The data for both countries were presented separately and annotated with cautions about the sample. The average percentage of questions correct for many of the participating countries were bunched around 60 per cent, including that for England. The large standard error for England's score contributes to this score not being significantly different from that of seven other participants including Spain, Ireland, Canada and the United States. Five countries, Korea, Hungary, Taiwan, the Soviet Union and Scotland outperformed England. The level of performance displayed by England

could be described as poor. England did not outperform any of the participating countries to a significant level.

England's performance in mathematics was not as good as that for science in the same survey, as discussed below.

This first view of comparative mathematics performance predates the implementation of the National Curriculum in England, but the next international survey in 1995 came after it had been established. This was under the IEA banner and known as TIMSS.

Mathematics: the TIMSS studies

The 1995 survey of mathematics was run by IEA, and was originally entitled the Third International Mathematics and Science Survey (TIMSS), (Mullis *et al.* 1997; Harris *et al.* 1997) but once the survey was established as the baseline for a series of such surveys it changed to the Trends in International Mathematics and Science Survey, thus maintaining the TIMSS acronym. This series of surveys is important in establishing England's performance level in two ways. On each occasion, as with the earlier surveys already discussed, a measure of the performance of England compared with other countries was given. Additionally, the TIMSS series of studies is linked by common items used in consecutive studies. This allows country performance in different surveys to be placed on the same scale, allowing within-country trends in performance to be identified.

To date there have been three TIMSS surveys, in 1995 (Mullis *et al.* 1997; Harris *et al.* 1997), in 1999 (Mullis *et al.* 2000) and in 2003 (Mullis *et al.* 2004; Ruddock *et al.* 2004). The 1999 survey did not include the primary age group, and so comparisons over time can only be based on the period from 1995 to 2003. In order to gain trend information from these two studies, they are discussed here as a pair.

Both surveys had similar structures, items being grouped into 'blocks' with each block appearing in several different tests. Each test included both mathematics and science item blocks, thus allowing each pupil to be given both a mathematics score on the mathematics scale and a science score.

Mathematics: TIMSS 1995

The 1995 TIMSS survey involved two adjacent cohorts, which in England were Years 4 and 5. The results discussed below are from the older group since that was the cohort also tested in later TIMSS surveys. The data from the younger cohort gave a very similar picture.

In 1995 the following countries outperformed England:

> Singapore, Korea, Japan, Hong Kong, Netherlands, Czech Republic, Austria, Slovenia, Ireland, Hungary, Australia, USA, Canada and Israel.

The countries generally performing at a similar level to England were:

> Latvia, Scotland, Cyprus, Norway and New Zealand.

The countries outperformed by England were:

> Greece, Thailand, Portugal, Iceland, Iran and Kuwait.

There are some similarities with the 1991 IAEP survey, in that Korea and Hungary again outperformed England. There were, however, several countries which had higher average scores than England in 1995 but had performed at a similar level in 1991 (USA, Canada and Ireland). Compared with 1991, England's performance was better against Scotland, which

had outperformed England in the earlier survey, and against Portugal, outscored by England in 1995, but not in 1991.

Science is discussed in more detail below, but the comparison of mathematics performance with science is illuminating. The relationship showed a strong similarity to the 1991 IAEP study; the relative standing of England in mathematics was not as high as in science. An illustration of this is that in science only three countries, Japan, Korea and the USA, outperformed England, while in mathematics 14 countries had higher levels of performance. At the other end of the performance spectrum, England outperformed 13 countries in science but only five in mathematics. None of the five countries outperformed by England in mathematics would be regarded as key economic competitors.

The next TIMSS survey to involve primary age pupils was in 2003, and the data from this survey allowed England's performance against other countries to be quantified and provided a direct measure of any change in England's performance over time.

Mathematics: TIMSS 2003 and trends over time

The mathematics data from TIMSS 1995 was rescaled together with that from 2003 to give scores on the same scale (Ruddock *et al.* 2004).

In 2003 the following countries had higher average scores than England:

Singapore, Hong Kong, Japan, Chinese Taipei, Belgium (Flemish) and Netherlands.

The countries generally performing at a similar level to England were:

Latvia, Lithuania, Russian Federation and Hungary.

The countries outperformed by England were:

United States, Italy, Australia, New Zealand, Scotland, Norway and eight other countries.

In general terms, the countries outperforming England in 2003 were from the Pacific Rim or Dutch-speaking Europe. On this occasion the countries with higher mean scores than England included several obvious economic competitors and benchmarks. The performance demonstrated by English students appeared to be much better than that in previous international surveys, and this can be explored in two ways: by looking at England's relative standing against other important comparison countries and by analysing England's scores over time.

Fifteen countries tested the same primary age group in the 1995 and 2003 TIMSS. England's performance level increased significantly from 1995 to 2003, rising from a scaled score of 484 to 531. This increase, 47 scale points, was the largest change in performance in any of the 15 countries participating in both 1995 and 2003. Six countries increased their performance in mathematics, seven showed no change and two showed a decline in performance.

It is also possible to look at trends over time via the common items, meaning those used in both the TIMSS surveys in 1995 and 2003. In grade 4 mathematics there were 37 such trend items. The average success rate for these items in England rose from 63 per cent to 72 per cent, a rise of 9 per cent. This shows a clear and marked increase in performance from 1995 to 2003 in primary mathematics performance in England. To put this in further context, Table 1 shows how England's trend in performance compares with that of a range of other participating countries.

Table 1: Trends in England's mathematics performance compared with other countries.

	1995	2003	Relative to other country, England performance	Other country's performance 1995 to 2003
Hong Kong				↑
Singapore				No change
Japan				No change
Netherlands				↓
Hungary			Improved	No change
United States		+	Improved	No change
Australia		+	Improved	No change
Scotland		+	Improved	No change
New Zealand		+	Improved	↑

+	England has higher level of performance than country shown.
	No significant difference between England and country shown
	England has lower level of performance than country shown.

England's performance improved against five of these countries, two of which, the United States and Australia, had outscored England in 1995. In none of these five countries can the improvement in England's relative standing be attributed to a decline in performance in the comparison country.

In summary, in primary mathematics the international surveys show performance in England to have been mediocre in the 1991 and 1995 surveys. England's performance improved considerably from 1995 to 2003. This improvement is clearly shown whether the change in England's score over this period is analysed or England's performance is compared with that of other participating countries. Nevertheless, the performance remains in the middle rank, below that of Pacific Rim and northern European countries, but significantly better than other English speaking countries such as the USA, Australia, New Zealand and Scotland. It would be hard not to attribute this change in mathematics performance to the influence of the National Curriculum in England from 1989 and the associated Numeracy Strategy in the late 1990s, both of which formalised the requirements on teachers and perhaps raised their expectations of pupils. However, there are other possible explanations and the international surveys cannot easily attribute causation to the differences they disclose and the changes they highlight.

Reading

In contrast to mathematics and science, the cycle of international surveys of literacy attainment has been sporadic. There were three IEA reading surveys in 1960 (Foshay *et al.* 1962), in 1971 (Thorndike 1973) and in 1991 (Elley 1992) and the written composition survey

of 1983 (Purves 1992). To that list can now be added the Progress in International Reading Literacy Study (PIRLS) of 2001 (Twist *et al.* 2003; Mullis *et al.* 2003a). The ages tested, as well as the number and nature of participating countries, has varied with each study.

The IEA survey in 1983 was the only international survey of writing attainment (Purves 1992). It involved 14 countries, including England and Wales (as one entity). The outcomes of this study differed from those of reading in that there was no comparative analysis of overall writing attainment between the countries, essentially due to the apparently insurmountable difficulties encountered in ensuring marking quality in writing in several different languages and from different education systems and curricula.

The 1960 IEA reading survey (Foshay *et al.* 1962) involved 12 countries, including England and Wales (participating jointly) and Scotland, and tested 13/14-year-olds. The 1971 IEA survey (Thorndike 1993) tested three age groups: 9-year olds, 13/14-year olds and 15/16-year olds. England and Wales again jointly participated in this survey, and Scotland was also represented, each at all three age ranges.

The 1991 IEA reading survey (Elley 1992) again involved 9-year olds. England was involved in the preparatory work for the study, including the pilot survey, but withdrew before the main survey took place. This was essentially because the model of reading being assessed was not thought by researchers at that time to adequately reflect the national curriculum, which was still a relatively recent innovation, or contemporary UK conceptions of the construct of reading:

> [the tests] 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.
>
> (Brooks *et al.* 1996, p. 3).

A study conducted in 1995, by Brooks, Pugh and Schagen (1996) provides some information about attainment at that point in relation to the attainment recorded in the 1991 survey, through the use of some of the IEA materials outside of the 'official' survey framework.

Results of international comparisons of reading (1960, 1971, 1991/96)

In the 1960 study of 13/14-year olds, England and Wales performed relatively well, and comfortably in the top half of a sample of 12 countries (Brooks 1997). Scotland's overall attainment was even better and was second only to the former Yugoslavia. In the 1971 study, at the age of 9/10 the mean reading attainment of pupils in England and Wales, and Scotland, was exceeded only by pupils in Sweden, Italy and Finland (Thorndike 1973). Pupils in a further eight countries, including those in the Netherlands and the United States, achieved less well. In this study, the standard deviation for England and Wales, used to measure the spread of scores, was equal highest (with the United States), indicating a very wide range of scores.

For the age 13/14 group, the performance of England and Wales was just below the median score for all participating countries, with Scotland being just above. England and Wales had the second highest standard deviation, after Israel, and Scotland had the third greatest spread, out of the total of 15 participating countries (Belgium represented twice, by French- and Flemish-speaking populations).

The results for the uppermost group being tested in this survey (15/16-year olds) can be contrasted with those of the two younger cohorts. England and Wales had the third highest mean score of all 15 participating countries and a standard deviation of just 0.1 above the median. The highest scoring country at this age group, as at age 14, was New Zealand.

Scotland was the second highest, and with a standard deviation below that of the median for all countries.

Perhaps the most interesting finding from the results for the three age groups assessed in 1971 is that, at least for the two younger groups, there is evidence for the wide range of attainment in England and Wales. This is one of the origins of the often asserted 'long tail of underachievement' in England. This phrase is used to describe the performance of less able pupils which is seen to 'tail off' dramatically and to lower the average score.

Following the introduction of the national curriculum in 1988 and its assessment in the early 1990s, attention turned once again to reading attainment in England relative to that of other countries. Brooks (1997) pointed out that there was evidence that reading attainment in England and Wales had been relatively stable in the years 1948-1979. A study which utilised components from the IEA 1991 survey and also a reading test (*Reading Ability Series*) which had been standardised in England and Wales in 1987 was conducted (Brooks *et al.* 1996). The researchers had a sample of 1,817 9-year old pupils in 58 schools and used a split design with each pupil taking one of the two main parts of the IEA survey instruments (and all pupils taking the vocabulary test) used in the original survey. Brooks *et al.* (1996) suggested that attainment in England and Wales in 1995 would have resulted in a position in about the middle of the international table in 1991. As the survey involved pupils with a mean age of 9 years 0 months, compared to the mean age of 9 years 8.4 months of pupils in the IEA survey, an age adjustment was made, following the procedure described in the IEA report (Elley 1992). This led to a slight rise in the overall standing for England and Wales, but this result remained within the middle grouping of countries.

One notable aspect of Brooks *et al.*'s study was the reaffirmation of the 'long tail of underachievement'. Using data from the 1995 follow up to the 1991 survey, the standard deviation for England and Wales was greater than that of 23 countries, equal to that of New Zealand, and smaller than that of three countries (Denmark, Norway and Sweden). Brooks *et al.* also stated that the phenomenon of the 'long tail' was seen not only in literacy studies, but also in international comparisons of mathematics and science.

The study of Brooks *et al*. (1996), while providing the only link with the IEA study conducted in 1991, nevertheless has some limitations, several of which are acknowledged in the published report. The range of scores achieved was narrower than the range achieved in the international survey. This was a function of the survey design and the authors suggest that it may have led to a ceiling effect, i.e. that some pupils could not show the full achievement of which they were capable. In addition to this, of the six open-ended questions in the IEA instruments, the two requiring a longer written response were not included in the analysis. This is relevant when the findings from PIRLS 2001 are considered below.

Progress in the International Reading Literacy Study (PIRLS 2001)

The results of the first of what is to be a five-yearly cycle of international comparisons of reading literacy, conducted under the auspices of the IEA, were published in 2003 and provided good evidence about England's reading standards in the 21st century (Twist *et al.* 2003; Mullis *et al.* 2003a).

Compared to earlier international surveys, great lengths were made to provide an explicit definition and framework for reading, within which the assessment instruments were conceived and the outcomes interpreted. The PIRLS definition of reading literacy is:

> The ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers, and for enjoyment.
>
> (Campbell *et al.* 2001).

	Purposes for reading		
Processes of comprehension	Literary experience	Acquire and use information	
Focus on and retrieve explicitly stated information			20%
Make straightforward inferences			30%
Interpret and integrate ideas and information			30%
Examine and evaluate content, language and textual elements			20%
	50%	50%	

In the PIRLS assessment framework (above), two central purposes for reading are identified: reading for literary experience, and reading to acquire and use information. Each purpose is characteristically associated with certain types of texts: reading for literary experience tends to be associated with the reading of stories or poems; reading to acquire and use information with factual texts such as instructional or informational texts.

On these two purposes for reading, the PIRLS framework superimposes four 'reading processes'. It is these processes which determine the type of questions which are asked about each of the texts.

Within the 35 participating countries, England's scale score in PIRLS 2001 was significantly lower than that of Sweden, not significantly different from the scale scores of the Netherlands and Bulgaria, and significantly higher than those of all other participating countries, including France, Germany, Italy, Scotland, New Zealand and the United States. It was therefore evidence of extremely high standards of reading in English primary schools for children at the age of about 10.

The PIRLS assessment scaled the scores of the participating countries on the two different reading purposes. On the literary experience scale, England and Sweden scored significantly higher than all the other 33 participating countries. Nine countries scored significantly higher than Scotland, and 18 countries scored significantly less well. The performance of Scotland was not significantly different to that of another seven countries.

When the scale of reading to acquire and use information is considered, a slightly different picture emerges. England's scale score was significantly lower than Sweden, was not significantly different from a further seven countries and was significantly higher than the remaining 26 countries. Scotland's scale score was significantly lower than 12 countries, including England, was not significantly different from those of a further seven countries, and was significantly lower than the remaining 15 countries.

When relative performance on the two scales of reading purposes is compared, England had one of the largest differences between the two scale scores (14 scale points). Scotland had a difference of two scale points. All of the countries that tested in English (England, New Zealand, Scotland, Singapore and the United States) did better on the scale measuring reading for literary purposes, although for two countries, Singapore and Scotland, the difference was small, at one and two scale points respectively. In contrast, some other countries, for example France, did much better on reading and using information than on literary reading, perhaps reflecting different cultural and curricular emphases.

A striking finding in PIRLS 2001 was that girls scored significantly higher than boys in all participating countries, echoing the finding of various assessments of reading, and English more widely, in England annually. This finding also held for the two purposes separately.

In addition to the high average achievement, the other most notable feature of the results from England was the wide range in achievement, also a feature of earlier surveys of reading. This is most readily described when the attainment of pupils at different points on the distribution is compared across countries. Table 2 below shows the scale score of pupils at the 5th, 25th, 50th, 75th and 95th percentiles for a subset of countries which participated in PIRLS 2001.

Table 2: Scale scores of pupils at the 5th, 25th, 50th, 75th and 95th percentiles in PIRLS

	5th percentile	25th percentile	50th percentile	75th percentile	95th percentile
England	395	501	559	612	685
France	403	481	528	573	636
Netherlands	458	517	556	593	645
New Zealand	360	472	537	593	668
Scotland	378	476	534	586	658
Sweden	445	521	565	605	663
United States	389	492	551	601	663

Adapted from Mullis *et al*. (2003a), Exhibit B.1

It is interesting to compare the range of scores from the Netherlands with those from England. Overall mean achievement was not significantly different in these two countries, but the pattern of performance across the ability range is very different. At the fifth percentile (i.e. where 95 per cent of pupils in the country scored higher), children in the Netherlands had the highest scale score of all 35 countries (458), and those in England at the fifth percentile had a scale score which was 15th highest (395). At the other end of the distribution, the highest achievers at the 95th percentile, the scale score for pupils in England (685) was the highest of all countries whereas the scale score of pupils in the Netherlands at the 95th percentile was bettered by pupils at this percentile in 11 other countries. The range between the 5th and 95th percentiles was 187 scale points for the Netherlands, the smallest in the study, and for England was 290 scale points, one of the largest.

As well as measuring reading attainment, data concerning various other aspects of reading, including pupils' attitudes, was collected as part of PIRLS by means of pupil questionnaires. Evidence of the attitudes of an earlier generation was collected by the Assessment of Performance Unit in the 1980s, which found that at least nine out of ten pupils indicated that they enjoyed reading stories, the strongest response to any of the attitude items, and there were also clear indications that the majority of pupils were positive about both reading independently and using books independently. Data collected for PIRLS 2001 gave a slightly less positive view of children's attitudes to reading but what raised greater concern was the fact that England had the second highest proportion of children who expressed clearly negative views about reading (13 per cent against an international average of 6 per cent). This was 18 per cent of boys in the sample and 8 per cent of girls. In the case of boys, just the Netherlands (23 per cent) and the United States (19 per cent) had a greater proportion in this 'low' category. Scotland came close behind England with 17 per cent. With respect to girls, the United States and England had jointly the greatest proportion of pupils expressing negative attitudes to reading, with Hungary, the Netherlands and Scotland in the next group (6 per cent).

Within all the participating countries, there was, unsurprisingly, a positive association between reading attainment and attitudes to reading. It is, though, interesting to note that that relationship did not exist between countries; the countries which had the highest overall attainment in PIRLS did not necessarily have the most positive attitudes to reading.

Science

International comparative surveys in science have been mounted from the 1970s, starting with the IEA First International Science Study in 1970-71, but this did not involve primary age pupils. However, the Second IEA International Science Study, administered in 1984, did involve this younger age group. England participated in this second study of science performance, and this study gave a first view of England's primary science performance compared with that of other countries (Postlethwaite and Wiley 1992).

Science: the 1984 IEA study

The target population for the study was all students aged 10 on the date of testing, or all students in the grade where most 10-year olds were to be found on the date of testing. In England, the definition was all pupils in Year 5 in the age range 10:0 to 10:11 at the start of the school year. The mean age of pupils tested in England was 10:3, somewhat younger than in most participating countries. Special schools were excluded from the study, and the response rate for schools was 66 per cent. This was similar to the response rates in Italy, Norway and Sweden but lower than those achieved in the Pacific Rim countries or in Eastern Europe.

The core test for the study consisted of 24 science items and was taken by each pupil. Each pupil also took two of a further four 8-item tests, giving 40 items per student from a total pool of only 56 items, many fewer than in later surveys. The content covered was classified as biology (22 items), physics (21), earth science (8) and chemistry (5).

The participants (for the primary population) in the 1984 study included 16 complete countries, plus Canada split into English and French speaking components and a second age cohort tested in Sweden. The presentation of the results for these early studies differed from that for later studies, where the statistical significance of differences in scores between countries are indicated, and it has been necessary to estimate which differences in performance between England and other participating countries are significant.

On all the available measures in the 1984 survey the following countries outperformed England:

Japan, Korea, Finland, Hungary, Italy, Australia, USA

The countries generally performing at a similar level to England were:

Singapore, Poland, Norway, Hong Kong

The countries outperformed by England were:

Philippines, Nigeria

Both Canadian language groups (English and French speaking) outperformed England, but the comparisons with Israel and the younger Swedish cohort were erratic. The older cohort in Sweden outperformed England on both measures.

The results of this survey do not suggest a high level of performance in science in England at that time. England did not, for example, outperform any of the developed countries in the survey. This set of data is important because it gives a picture of comparative performance by English pupils before the National Curriculum was introduced.

Science: the 1991 IAEP study

The next international science survey took place in 1988, organized by the IAEP, but did not involve primary age pupils. A further study was carried out by the same organisation in 1991, and this time primary students were involved (Lapointe *et al.* 1992a). The target age group was nine-year olds, and pupils from England were drawn from the two year groups containing pupils of this age.

The response rate for schools in England was 56 per cent, the lowest of all the participating countries, but not much worse than Scotland. Again, estimates of the significances of the differences between countries have had to be made since these studies did not calculate them. The average percentage correct scores for many of the participating countries were bunched around 62 per cent. Treating the results with caution, it is reasonable to conclude that Korea and Taiwan outperformed England, while England outperformed Slovenia, Ireland and Portugal. England's results were similar to those for the USA, Canada, Hungary, Scotland, Spain, the Soviet Union and Israel.

Comparisons with the previous survey are hampered by the relative scarcity of countries participating on both occasions. Korea clearly outperformed England on both occasions. The USA and Canada had outperformed England in the 1984 IEA survey, but performed at a similar level in 1991. This survey took place during the initial stages of implementing the National Curriculum, noticeably so for the age group tested, but the next international survey, in 1995 does represent England's performance when the National Curriculum had just been established.

Science: the TIMSS studies

As explained above, the 1995 IEA survey included both maths and science and was originally entitled the Third International Mathematics and Science Survey but once the survey was established as the baseline for a series of such surveys it changed to the Trends in International Mathematics and Science Survey (TIMSS). In each subsequent survey, a snapshot of the performance of England compared with other countries was given and the country's performance in different surveys were placed on the same scale, allowing within-country trends in performance to be identified.

Science: TIMSS 1995

The TIMSS 1995 survey involved two adjacent cohorts, and in England these were Years 4 and 5. The science results discussed below are from the older group since that was the cohort also tested in later TIMSS surveys. In fact, the data from the younger cohort gave a very similar picture (Martin *et al*. 1997; Harris *et al.* 1997).

In 1995 the following countries only outperformed England:

> Japan, Korea and USA.

The countries generally performing at a similar level to England were:

> Austria, Australia, Netherlands, Czech Republic, Canada, Singapore, Slovenia, Ireland and Scotland.

The countries outperformed by England were:

> Hong Kong, Hungary, New Zealand, Norway, Latvia, Israel, Iceland, Greece, Portugal, Cyprus, Thailand, Iran and Kuwait

The picture of England's science performance obtained in 1995 was rather different from that shown in the earlier surveys. Only three countries outperformed England, two from the Pacific Rim and the USA. England performed at a similar level to several of its European neighbours, including Scotland and Ireland, and outperformed four others. Overall, the level of performance demonstrated by English students was high.

Looking back to the earlier studies, a number of common patterns can be identified in England's performance relative to other countries up to 1995. Japan and Korea consistently outperformed England, while the USA, Canada and Australia performed at a level higher than or similar to England. Singapore and Scotland performed at a similar level to England. It should be noted that changes in England's performance relative to other countries could have been caused by a change in performance in England, a change in performance in the country being compared with England, or a combination of the two. The TIMSS 2003 data allows judgments on which of these factors are involved and is discussed below.

Science: TIMSS 2003 and trends over time

The next full TIMSS survey was in 2003 (Martin *et al.* 2004b; Ruddock *et al.* 2004). The science data from TIMSS 1995 was rescaled together with that from 2003 to give scores on the same scale.

In 2003, only Singapore and Chinese Taipei outperformed England, with Japan, Hong Kong and the USA performing at a similar level. England's score was significantly higher than that of all the other participating countries. Again England showed a high level of performance, outscoring all the other European countries which participated.

Fifteen countries tested the same primary age group in the 1995 and 2003 TIMSS studies. England's performance level increased significantly from 1995 to 2003, rising from a scaled score of 528 to 540. Of the 15 countries, nine increased their performance, three showed no change and three showed a decline in performance. Most of the countries showing an increase in score from 1995 to 2003 had scores lower than England's in 1995. The increase in England's score, 13, was one of the smaller increases which occurred; large increases were made by, for example, Singapore (42) and Hong Kong (35). Norway showed the largest decline (38 scale points).

It is also possible to look at trends over time via the items used in both TIMSS 1995 and 2003. In grade 4 science there were 32 such items. The average success rate for these items in England rose by 4 per cent from 76 per cent to 80 per cent.

Table 3 examines England's change in performance relative to a range of other participants. England's performance improved against five of these countries, two of which, Japan and Scotland, had lower scores in 2003 than in 1995. In spite of England's improved performance, ground was lost against both Singapore and Hong Kong, countries with larger increases in score than England over this period.

Table 3: Trends in England's science performance compared to other countries.

	1995	2003	Relative to other country, England's performance	Other country's performance 1995 to 2003
Japan			Improved	↓
United States			Improved	No change
Netherlands		+	Improved	No change
Australia		+	Improved	No change
Scotland		+	Improved	↓
Hungary	+	+		↑
New Zealand	+	+		↑
Hong Kong	+		Declined	↑
Singapore			Declined	↑

+	England has higher level of performance than country shown.
	No significant difference between England and country shown.
	England has lower level of performance than the country shown.

In summary, the international surveys provide clear evidence of a rise in Year 5 performance for science from 1995 to 2003. The 1995 level of performance was already high, amongst the highest in the participating countries, and this good performance in primary science has continued. Before 1995 it is more difficult to make comparisons with other countries. The available data is sparse and few countries participated in several of the surveys undertaken. It does, however, seem that England's performance in science in the surveys carried out before 1995 was not outstanding.

Conclusion

Direct evidence on the performance of primary school pupils in England from international surveys is sparser than might be expected. Prior to the 1990s, international surveys were irregular and methodologically weak. The number which included primary children was rather small. Recently, international organisations have established regular cycles of surveys which give the prospect of better examinations of trends of time. One series, the OECD's PISA, has thus far concentrated only on the outcomes of schooling and not directly addressed primary children. The other series, that of the IEA, has addressed the attainment of primary school children in mathematics, science and reading.

The available evidence is that the level omathematics performance is currently in the middle rank, below that of Pacific Rim and northern European countries, but significantly better than some other English speaking countries such as the USA, Australia, New Zealand and Scotland. This middle ranking does though represent a slight improvement from earlier surveys in which England's performance was very poor.

There are greater cultural problems with the assessment for reading, and fewer surveys. The most recent survey, PIRLS in 2001, indicated that the reading skills of English pupils were among the highest in the world, with good achievement in both literary and information reading. This does seem to have been an improvement on the standing in earlier surveys, though the reliability of the evidence from those is weak. There is some evidence that this high attainment is at the expense of enjoyment of reading. The 2006 PIRLS survey will report in November 2007, giving information on trends in reading performance over time.

Primary science represents something of a success story for England. There is clear evidence of a rise in performance from 1995 to 2003 even though England was amongst the highest in the participating countries in the 1990s. Before 1995 the available data is sparse but it does seem that England's performance in science in earlier surveys was at a lower level.

A consistent factor in England's results across all three subject areas is a high range of scores, compared to many other countries. High attaining English pupils are among the top ranking in the world in reading and science, but the greater spread of attainment means that the low attaining pupils are far below these in their attainment. For mathematics, the average performance is also poor by the standards of other English speaking countries and those of many European and international competitor countries.

International surveys now have a robust but not perfect methodology and are an important source of information on the relative performance of England's education system. Since their data is publicly available, they are also a resource for much secondary analysis, as yet relatively unused. There are further studies in progress, PIRLS 2006 and TIMSS 2007, which will continue the time series of comparative data.

References

Beaton, A.E., Postlethwaite, T.N., Ross, K.N., Spearritt, D. and Wolf, R.M. (1999). *The Benefits and Limitations of International Education Achievement Studies.* Paris: International Institute for Educational Planning/UNESCO.

Bechger, T.M., van Schooten, E., De Glopper, C. and Hox-Joop, J.J. (1998). 'The validity of international surveys of reading literacy: the case of the IEA Reading Literacy Study', *Studies in Educational Evaluation*, **24**, 2, 99–125.

Blum, A., Goldstein, H. and Guerin-Pace, F. (2001). 'International Adult Literacy Survey (IALS): an analysis of international comparisons of adult literacy', *Assessment in Education*, **9**, 3, 388–399.

Bonnet, G. (2002). 'Reflections in a Critical Eye: on the pitfalls of international assessment', *Assessment in Education*, **9**, 3, 387–399.

Bonnet, G., Braxmeyer, N., Horner, S., Hannu-Pekka L., Levasseur, J., Nardi, E., Remond, M., Vrignaud, P. and White, J. (2001). *The Use of National Reading Tests for International Comparisons: Ways of Overcoming Cultural Bias*. Paris: Ministère de l'Education Nationale, Direction de la Programmation et du Dévelopment.

Bonnet, G., Daems, F., de Glopper, C., Horner, S., Lappalainen, H-P., Nardi, E., Remond, M., Robin, I., Rosen, M., Solheim, R.G., Tonnessen, F-E., Vertecchi, B., Vrignaud, P., Wagner, A.K.H. and White, J. (2003). *Culturally Balanced Assessment of Reading (c-bar). A European Project* [online]. Available: http://cisad.adc.education.fr/reva/pdf/cbarfinalreport.pdf [6 August, 2007].

Brooks, G. (1997). 'Trends in standards of literacy in the United Kingdom, 1948-1996.' Paper presented at British Educational Research Association conference, University of York, September [online]. Available: http://www.leeds.ac.uk/educol/documents/000000650.htm [6 August, 2007].

Brooks, G., Foxman, D. and Gorman, T.P. (1995). *Standards in Literacy and Numeracy: 1948-1994* (NCE Briefing New Series 7). London: National Commission on Education.

Brooks, G., Pugh, A. and Schagen, I. (1996). *Reading Performance at Nine*. Slough: NFER.

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O. and Sainsbury, M. (2001). *Framework and Specifications for PIRLS Assessment 2001.* Second edn. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center [online]. Available: http://timss.bc.edu/pirls2001i/pdf/PIRLS_frame2.pdf [6 August, 2007].

Clark, M.M. (2004). 'International studies of reading, such as PIRLS – a cautionary tale', *Education Journal*, **75**, 25–27.

Elley, W.B. (1992). *How in the World do Students Read? IEA Study of Reading Literacy.* The Hague: International Association for the Evaluation of Educational Achievement.

Foshay, A.W., Thorndike, R.L., Hotyat, F., Pidgeon, D.A. and Walker, D.W. (1962). *Educational Achievements of Thirteen Year Olds in Twelve Countries: Results of an International Research Project, 1959-61*. Hamburg: UNESCO Institute for Children.

Galton, M. (1998). 'What do the tests measure?' *Education 3-13*, **26**, 2, 50–59.

Goldstein, H. (2004). 'International comparative assessment: how far have we really come?' (Review Essay), *Assessment in Education*, **11**, 2, 227–234.

Harris, S., Keys, W. and Fernandes. C. (1997). *Third International Mathematics and Science Study, Second National Report. Part 1: Achievement in Mathematics and Science at Age 9 in England.* Slough: NFER.

Hilton, M. (2006). 'Measuring standards in primary English: issues of validity and accountability with respect to PIRLS and National Curriculum test scores', *British Educational Research Journal*, **32**, 6, 817–837.

Husén, T. and Tuijnman, A. (1994). 'Monitoring standards in education: why and how it came about.' In: Tuijnman, A. and Postlethwate, T.N. *Monitoring the Standards of Education*. Oxford: Elsevier Science.

Joncas, M. (2003). 'PIRLS sampling weights and participation rates.' In: Martin, M.O., Mullis, I.V.S. and Kennedy, A.M. *PIRLS 2001 Technical Report*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Kelly, D.L. and Malak, B. (2003). 'Translating the PIRLS reading assessment and questionnaires.' In: Martin, M.O., Mullis, I.V.S. and Kennedy, A.M. *PIRLS 2001 Technical Report*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Lafontaine, D. (2004). 'From comprehension to literacy: thirty years of reading assessment.' In: Moskowitz, J.H. and Stephens, M. *Comparing Learning Outcomes: International Assessment and Education Policy*. London: RoutledgeFalmer.

Lafontaine, D. and Demeuse, D. (2002). 'Le bon (critique), la brute (médiatique) et les truands (anglo-saxons)', *Revue Nouvelle*, **3-4**, 115, 100–108.

Lapointe, A.E., Askew, J.M. and Mead, M.A. (1992a). *Learning Science.* Princeton, NJ: Educational Testing Service.

Lapointe, A.E., Mead, M.A and Askew, J.M. (1992b). *Learning Mathematics.* Princeton, NJ: Educational Testing Service.

Martin, M.O., Mullis, I.V.S., Beaton, A.E., Gonzalez, E.J., Smith, T.A. and Kelly D.L. (1997). *Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Martin, M.O., Mullis, I.V.S. and Kennedy, A.M. (2003). *PIRLS 2001 Technical Report.* Chestnut Hill, MA: Boston College, International Study Center [online]. Available: http://timss.bc.edu/pirls2001i/PIRLS2001_Pubs_TR.html [7 August, 2007].

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J. and Chrostowski, S.J. (2004a). *TIMSS 2003 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Martin, M.O., Mullis, I.V.S. and Chrostowski, S.J. (Eds) (2004b). *TIMSS 2003 Technical Report. Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O and Sainsbury, M. (2006). *PIRLS 2006 Assessment Framework and Specifications.* Second edn. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.O., Kelly, D.L. and Smith, T.A. (1997). *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: Boston College, Centre for the Testing, Evaluation, and Educational Policy.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.O. and Chrostowski, S.J. (2004). *TIMSS 2003 International Mathematics Report: Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J. and Smith T.A. (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College, International Study Center.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J. and Kennedy, A.M. (2003a). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary Schools in 35 Countries.* Chestnut Hill, MA: Boston College, International Study Center.

Mullis, I.V.S., Martin, M.O. and Kennedy, A.M. (2003b). 'Item analysis and review.' In: Martin, M.O., Mullis, I.V.S. and Kennedy, A.M. *PIRLS 2001 Technical Report*. Chestnut Hill, MA: Boston College, International Study Center [online]. Available: http://timss.bc.edu/pirls2001i/PIRLS2001_Pubs_TR.html [7 August, 2007].

Mullis, I.V.S., Martin, M.O., Ruddock, G. J., O'Sullivan, C.Y., Arora, A. and Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks.* Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Organisation for Economic Co-operation and Development (2004). *Learning for Tomorrow's World: First Results from PISA 2003* [online]. Available: http://www.oecd.org/dataoecd/1/60/34002216.pdf [2 August, 2007].

Organisation for Economic Co-operation and Development (2006). *Assessing Scientific, Reading and Mathematical Literacy: a Framework for PISA 2006.* Paris: OECD.

Postlethwaite, T.N. and Wiley, D.E. (1992). *The IEA Study of Science II: Science Achievement in Twenty-Three Countries.* Oxford: Pergamon.

Purves, A.C. (1973). *Literature Education in Ten Countries: an Empirical Study* (IEA International Studies in Education 2). Stockholm: Almqvist and Wiksell.

Reynolds, D. and Farrell, S. (1996). *Worlds Apart? A Review of International Surveys of Educational Achievement.* London: OFSTED.

Ruddock, G., Sturman, L., Schagen, I., Styles, B., Gnaldi, M. and Vappula, H. (2004). *Where England Stands in the Trends in International Mathematics and Science Study (TIMSS) 2003: National Report for England*. Slough: NFER.

Smithers, A. (2004). *England's Education: What Can be Learned by Comparing Countries?* Liverpool: Centre for Education and Employment Research.

Thorndike, R.L. (1973). *Reading Comprehension Education in Fifteen Countries* (IEA International Studies in Education 3). Stockholm: Almqvist and Wiksell.

Twist, L., Sainsbury, M., Woodthorpe, A. and Whetton, C. (2003). *Reading All Over the World: Progress in International Reading Literacy Study PIRLS. National Report for England.* Slough: NFER.

Van der Linden, W.J. and Hambleton, R.K. (Eds) (1997). *Handbook of Modern Item Response Theory.* New York, NY: Springer.

Whetton, C., Twist, L. and Sainsbury, M. (2007, forthcoming). 'Measuring standards in primary English: the validity of PIRLS: a response to Mary Hilton', *British Educational Research Journal*, **33**, 6.

Winter, S. (1998). 'International comparisons of student achievement', *Education 3-13*, **26**, 2, 26–33.

APPENDIX 1

THE PRIMARY REVIEW PERSPECTIVES, THEMES AND SUB THEMES

The Primary Review's enquiries are framed by three broad perspectives, the third of which, primary education, breaks down into ten themes and 23 sub-themes. Each of the latter then generates a number of questions. The full framework of review perspectives, themes and questions is at www.primaryreview.org.uk

The Review Perspectives

P1 Children and childhood
P2 Culture, society and the global context
P3 Primary education

The Review Themes and Sub-themes

T1 Purposes and values
- T1a Values, beliefs and principles
- T1b Aims

T2 Learning and teaching
- T2a Children's development and learning
- T2b Teaching

T3 Curriculum and assessment
- T3a Curriculum
- T3b Assessment

T4 Quality and standards
- T4a Standards
- T4b Quality assurance and inspection

T5 Diversity and inclusion
- T5a Culture, gender, race, faith
- T5b Special educational needs

T6 Settings and professionals
- T6a Buildings and resources
- T6b Teacher supply, training, deployment & development
- T6c Other professionals
- T6d School organisation, management & leadership
- T6e School culture and ethos

T7 Parenting, caring and educating
- T7a Parents and carers
- T7b Home and school

T8 Beyond the school
- T8a Children's lives beyond the school
- T8b Schools and other agencies

T9 Structures and phases
- T9a Within-school structures, stages, classes & groups
- T9b System-level structures, phases & transitions

T10 Funding and governance
- T10a Funding
- T10b Governance

APPENDIX 2

THE EVIDENTIAL BASIS OF THE PRIMARY REVIEW

The Review has four evidential strands. These seek to balance opinion seeking with empirical data; non-interactive expressions of opinion with face-to-face discussion; official data with independent research; and material from England with that from other parts of the UK and from international sources. This enquiry, unlike some of its predecessors, looks outwards from primary schools to the wider society, and makes full though judicious use of international data and ideas from other countries.

Submissions

Following the convention in enquiries of this kind, submissions have been invited from all who wish to contribute. By June 2007, nearly 550 submissions had been received and more were arriving daily. The submissions range from brief single-issue expressions of opinion to substantial documents covering several or all of the themes and comprising both detailed evidence and recommendations for the future. A report on the submissions will be published in late 2007.

Soundings

This strand has two parts. The *Community Soundings* are a series of nine regionally based one to two day events, each comprising a sequence of meetings with representatives from schools and the communities they serve. The Community Soundings took place between January and March 2007, and entailed 87 witness sessions with groups of pupils, parents, governors, teachers, teaching assistants and heads, and with educational and community representatives from the areas in which the soundings took place. In all, there were over 700 witnesses. The *National Soundings* are a programme of more formal meetings with national organisations both inside and outside education. They will take place during autumn 2007 and will explore key issues arising from the full range of data thus far. They will aim to help the team to clarify matters which are particularly problematic or contested and to confirm the direction to be taken by the final report. As a subset of the National Soundings, a group of practitioners - the *Visionary and Innovative Practice (VIP) group* – is giving particular attention to the implications of the emerging evidence for the work of primary schools.

Surveys

30 surveys of published research relating to the Review's ten themes have been commissioned from 69 academic consultants in universities in Britain and other countries. The surveys relate closely to the ten Review themes and the complete list appears in Appendix 3. Taken together, they will provide the most comprehensive review of research relating to primary education yet undertaken. They will be published in thematic groups from October 2007 onwards.

Searches

With the co-operation of DfES/DCSF, QCA, Ofsted, TDA and OECD, the Review is re-assessing a range of official data bearing on the primary phase. This will provide the necessary demographic, financial and statistical background to the Review and an important resource for its later consideration of policy options.

Other meetings

In addition to the formal evidence-gathering procedures, the Review team meets members of various national bodies for the exchange of information and ideas: government and opposition representatives; officials at DfES/DCSF, QCA, Ofsted, TDA, GTC, NCSL and IRU; representatives of the teaching unions; and umbrella groups representing organisations involved in early years, primary education and teacher education. The first of three sessions with the House of Commons Education and Skills Committee took place in March 2007. Following the replacment of DfES by two separate departments, DCSF and DIUS, it is anticipated that there will be further meetings with this committee's successor.

APPENDIX 3

THE PRIMARY REVIEW INTERIM REPORTS

The interim reports, which will be released in stages from October 2007, include the 30 research surveys commissioned from external consultants together with reports on the community soundings and the submissions prepared by the Cambridge team. They are listed by Review theme below, although this will not be the order of their publication. Report titles may be subject to minor amendment.

Once published, the interim reports, together with briefings summarising their findings, may be downloaded from the Review website, www.primaryreview.org.uk .

1. *Community Soundings: report on the Primary Review regional witness sessions*
2. *Submissions received by the Primary Review*
3. *Aims and values in primary education*. Research survey 1/1 (John White)
4. *The aims of primary education: England and other countries*. Research survey 1/2 (Maha Shuayb and Sharon O'Donnell)
5. *The changing national context of primary education*. Research survey 1/3 (Stephen Machin and Sandra McNally)
6. *The changing global context of primary education*. Research survey 1/4 (Hugh Lauder, John Lowe and Dr Rita Chawla-Duggan)
7. *Children in primary schools: cognitive development*. Research survey 2/1a (Usha Goswami and Peter Bryant)
8. *Children in primary schools: social development and learning*. Research survey 2/1b (Christine Howe and Neil Mercer)
9. *Teaching in primary schools*. Research survey 2/2 (Robin Alexander and Maurice Galton)
10. *Learning and teaching in primary schools: the curriculum dimension*. Research survey 2/3 (Bob McCormick and Bob Moon)
11. *Learning and teaching in primary schools: evidence from TLRP*. Research survey 2/4 (Mary James and Andrew Pollard)
12. *Curriculum and assessment policy: England and other countries*. Research survey 3/1 (Kathy Hall and Kamil Øzerk)
13. *The impact of national reform: recent government initiatives in English primary education*. Research survey 3/2 (Dominic Wyse, Elaine McCreery and Harry Torrance)
14. *Curriculum alternatives for primary education.* Research survey 3/3 (James Conroy and Ian Menter)
15. *The quality of learning: assessment alternatives for primary education*. Research survey 3/4 (Wynne Harlen)
16. *Standards and quality in English primary schools over time: the national evidence*. Research survey 4/1 (Peter Tymms and Christine Merrell)
17. *Standards in English primary schools: the international evidence*. Research survey 4/2 (Chris Whetton, Graham Ruddock and Liz Twist).
18. *Quality assurance in primary education*. Research survey 4/1 (Peter Cunningham and Philip Raymont)
19. *Children, identity, diversity and inclusion in primary education.* Research survey 5/1 (Mel Ainscow, Alan Dyson and Jean Conteh)
20. *Children of primary school age with special needs: identification and provision*. Research survey 5/2 (Harry Daniels and Jill Porter)

21. *Children and their primary education: pupil voice*. Research survey 5/3 (Carol Robinson and Michael Fielding)

22. *Primary education: the physical environment*. Research survey 6/1 (Karl Wall, Julie Dockrell and Nick Peacey)

23. *Primary education: the professional environment*. Research survey 6/2 (Ian Stronach, Andy Pickard and Elizabeth Jones)

24. *Teachers and other professionals: training, induction and development*. Research survey 6/3 (Olwen McNamara, Rosemary Webb and Mark Brundrett)

25. *Teachers and other professionals: workforce management and reform.* Research survey 6/4 (Hilary Burgess)

26. *Parenting, caring and educating*. Research survey 7/1 (Yolande Muschamp, Felicity Wikeley, Tess Ridge and Maria Balarin)

27. *Children's lives outside school and their educational impact*. Research survey 8/1 (Berry Mayall)

28. *Primary schools and other agencies*. Research survey 8/2 (Ian Barron, Rachel Holmes, Maggie MacLure and Katherine Runswick-Cole)

29. *The structure and phasing of primary education: England and other countries*. Research survey 9/1 (Anna Eames and Caroline Sharp)

30. *Organising learning and teaching in primary schools: structure, grouping and transition*. Research survey 9/2 (Peter Blatchford, Judith Ireson, Susan Hallam, Peter Kutnick and Andrea Creech)

31. *The financing of primary education.* Research survey 10/1 (Philip Noden and Anne West)

32. *The governance, administration and control of primary education*. Research survey 10/2 (Maria Balarin and Hugh Lauder)

The Primary Review is a wide-ranging independent enquiry into the condition and future of primary education in England. It is supported by Esmée Fairbairn Foundation, based at the University of Cambridge and directed by Robin Alexander. The Review was launched in October 2006 and aims to publish its final report in autumn 2008.

FURTHER INFORMATION

www.primaryreview.org.uk

General enquiries: enquiries@primaryreview.org.uk

Media enquiries: richard@margrave.co.uk

Published by the Primary Review,
Faculty of Education, University of Cambridge
184 Hills Road, Cambridge, CB2 8PQ, UK

ISBN 978-1-906478-02-5