A comparative investigation of England's national assessments and the international surveys

May 2010
Author: Fatima Lampreia Carvalho
Ofqual/10/4710

Contents

Introduction

The purpose of the research work underpinning this report was to answer a simple question: if validity must always be the grounds for the inferences drawn from assessment records (Cronbach, 1989; Embretson, 1983; Kane, 1992; Messick 1989, 1994), how can the international assessments provide more evidence on the validity of the national assessments? This report was originally written for the Ofqual audience who could be helped by a study introducing the international assessments as potential additional checks on the standards of the national curriculum assessments, the General Certificate of Secondary Education (GCSEs) and other qualifications. However, the present format also seeks to engage the wider assessment community in a discussion on the possibility of using the international assessments as external reference on the performance of pupils sitting both national and international assessments. It is in Ofqual's interest to generate debate and reach consensus on questions such as the following. Even if there is no curriculum match between the underlying focuses of the international assessments and the theories, processes and cognitive demands specified in the national curriculum assessments and GCSEs, can one use the international assessments to validate statutory national assessments? Does it make sense to tap into the validity of national assessments from a concurrent validity[1] point of view? Should Ofqual carry out regular checks on how well performance in the national curriculum assessments and GCSEs matches evidence already available from the international assessments?

Validation exercises require the accumulation of empirical data and logical arguments to guarantee the strength of inferences made from assessment results. The notion of validity had traditionally encompassed three broad categories: content validity, criterion-related validity and construct validity. However, considering validation in educational measurement, Samuel Messick stressed the concept of construct validity as a 'unitary principle' (Messick, 1989, 1996a, 1996b) working as the main reference for weighting up the uses and consequences of assessments.. Accepting the view of Cyril Weir (2005) that construct validity is a multifaceted concept linking content, domain processes, scoring structure, coverage, external factors and consequential aspects, this report concentrates on *concurrent validity*, which refers to the validation of score values by means of comparison of assessments with other assessments.[2]

It seems that a strong approach to concurrent validation would require the direct analysis of national and international assessments samples from the same

[1] For an introductory approach to criterion-related (concurrent and predictive) validity, see http://www.nfer.ac.uk/nfer/research/assessment/about/validity.cfm (accessed 6 February 2010). See also Wiliam, D. (2009).

[2] Weir, C.J. (2005), pp. 43–9.

candidates' scripts to demonstrate whether the level of conceptual understanding on a topic in the national assessments is indeed related to the level of understanding on the same topic in the international assessments. Concurrent validity studies are slightly different from predictive approaches, which would ask how well high international assessment performance predicts a future criterion – for example high GCSE grades or high performance in GCE A level examinations – that is used as an extra reference for university selectors. In this case an international score would have to predict school and university performance with reasonable accuracy. Nonetheless, it would need a number of years to set up predictive validity. Given the volatile context of regulatory policy making, it seems that concurrent validation of the national assessments and examinations can be the most promising to the regulatory remit of safeguarding the validity of the national assessments and examinations.

Concurrent validation exercises should first seek evidence that the assessment design supporting the national and international assessments intend to measure the same construct. American psychologist Samuel Messick (1994, p. 17) theorised that a construct-centred approach to assessment design should concentrate on what composite of knowledge, skills and further attributes ought to be assessed. Assessments should tie the knowledge and skills to be assessed to the explicit or implicit objectives of instruction Construct validity checks should also aim at what kind of performances can reveal the intended constructs, and what tasks or situations should elicit those behaviours. A study on concurrent validity between different assessments needs therefore to build on knowledge of the construct underpinning such assessment design.

This report initiates a study of the knowledge, skills and further attributes assessed by England's national assessments and GCSEs in relation to some international assessments undertaken by samples of England's students – the long-established series of studies Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS) and the newcomer Programme for International Student Assessment (PISA). TIMSS (since 1994–5) and PIRLS are conducted by the International Association for the Evaluation of Educational Achievement (IEA) – an independent international cooperative of national research institutions and governmental research agencies. PISA has been conducted by the intergovernmental Organisation for Economic Cooperation and Development (OECD) since 2000.[3]

[3] For a sound approach to methodological criticism made of the international surveys, view 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, pp. 3–7, http://www.primaryreview.org.uk/Downloads/Int_Reps/2.Standards_quality_assessment/Primary_Revi

Our goal may seem ambitious for a desk-based research study, but the reader should expect only a systematic comparison of the key features of national and international assessments regarding assessment features, content specifications, measurement scales and curriculum match between processes and cognitive demands in the diverse assessments.

Background

This report builds on two research outputs aiming to equip the Ofqual Assessment Research Team with broad expertise in international assessing, one being an extensive comparison between the national and international assessments, and the other a short document revealing what a small group of key interviewees think about the international assessments. From an editorial perspective, the appraisals of the international assessments did not harmonise with the comparative investigation of England's national assessments and international assessments. Some interviewees touched on policy making, others discussed curricular overlap and the cultural limitations of international assessments, citing a key study carried by Bonnet (2002), while some researchers recommended more funding so that the national assessment community can tap into the international background questionnaires database. The raw materials underpinning this short document were interviews conducted in 2008/09 with professionals from the Department for Children, Schools and Families (DCSF), Ofqual, QCDA and the National Foundation for Educational Research (NFER).

The general overview of the international assessing project also aimed to provide an up-to-date briefing paper on key issues for Ofqual staff in terms of communication policy. Some media coverage of the national and international trends on educational achievement could be misleading, undermining public understanding. The regulator would need to avoid and counteract simplistic media approaches to assessment results. An ideal communication policy would be to supply the media with an explanation for the rise (and fall) in the standards of attainment of English candidates, and if the regulator has an in-depth understanding of how to compare the international assessments (TIMSS, PIRLS and PISA), the national curriculum assessments and GCSEs, such objectives would be greatly facilitated.

Provided that Ofqual acquired a good understanding of the international assessments, interviewees working on 14–19 regulation envisaged a good use for TIMSS, PIRLS and PISA results over time. However, there are four reasons why international assessments may not be a reliable system of checks and balances to help to validate and monitor the national curriculum assessments and GCSEs. First, the international assessments are infrequent. PIRLS happens only every five years,

---

ew_WhettonRuddockTwist_4-2_briefing_Standards_-_International_evidence_071102.pdf (Last consulted 16 February, 2010)

and TIMSS takes place every four years. PISA is the most frequent survey, but still there is a three-year lag between assessments, with a focus on a major subject happening every nine years. No checks of the national assessments would be possible in between the international cycles. Second, there is only a relatively small sample of respondents to the international assessments. Once one gets down to comparing between specific parts of subjects, one might end up with a small data set for each curriculum focus. Third, the international assessments have a higher profile in countries like Norway and Germany. If PIRLS, TIMSS and PISA were subject to greater scrutiny in England, more questions about them would be raised. Fourth, the international assessments may work better for validating national curriculum assessments, which address broader constructs than do GCSEs in individual subjects. GCSEs can be more detailed in their own curriculum demands than a general international assessment may be. Concurrent validation of GCSE biology in relation to PISA science may be hard to sustain.

Given the number of technical issues that one would need to overcome before proposing any systematic comparisons between national and international assessments, this report only proposes a reasonable starting point for thinking about how to use the international assessments. In order to deliver this outline, we describe the overlapping areas between the national and international assessments, explaining whether the constructs being tested by a particular pair of assessments are the same. There are doubts about whether TIMSS mathematics and science test pupils from England on the same curriculum assessed by the national curriculum assessments. It is not clear whether there is any resemblance between the curriculum assessed by English, mathematics and science GCSEs and the content assessed by PISA. Comparing technical aspects of the national and international assessments is a massive but much required task.

There are some political reasons for using the international assessments to keep the national system in check. The international assessments aree deemed to be 'disinterested', with no accountability issues attached to them. The national curriculum assessments and the GCSEs, on the other hand, are part of the English accountability regime, and therefore there is much national and political interest in them. However, if one made the international assessments – in effect – part of the English accountability regime, they would become controversial too. Therefore using the international assessments as an option for monitoring performance in the national assessments would make sense only if these comparisons were informal exercises to gain insight into performance under different assessment conditions.

An initial written report completed in June 2009 highlighted major issues in international assessing, such as assessment features, content specifications, quality assurance, measurement scales, reporting, curriculum match and trends in achievement. Such issues were not discussed in great detail as the original project aimed at breadth rather than depth, to survey the area with appropriate reference to

where detail could be found in the literature. This paper is a summary of the original work targeting a broader audience, especially the national and international assessment community, academics working in the field of education and assessment, awarding body professionals, and governmental and non-governmental bodies that wish to participate in a dialogue on how individual nations can benefit from the international assessments to promote the quality of their internal assessments. Therefore the significance and potential contribution of this study is to engage the Ofqual audience and the wider assessment community in a quest for the best uses of the international assessments as external references that safeguard the validity of national assessments.

Methodology

This small qualitative inquiry used exploratory, explanatory and descriptive strategies to investigate the connections between the national and international assessments. By consulting key stakeholders from the DCSF, NFER and QCDA, and also assessment experts abroad (USA, Israel, Holland), the lead researcher carried out in-depth interviews that helped to identify the kinds of issues that would be useful for consideration by Ofqual colleagues in policy positions. Data collection was organised in six categories:

1. International technical reports covering PIRLS 2006, TIMSS 2007 and PISA 2006, with attention to assessment framework, specifications and achievement.

2. National reports on TIMSS, PIRLS and PISA commissioned by the DSCF.

3. National reports produced by the NFER, and reports published by the International Unit at QCDA and disseminated through the International Review of Curriculum and Assessment Frameworks Internet Archives (INCA). [4]

4. England's regulatory documents published by Ofqual and QCDA, such as the framework for the national curriculum assessments, programmes of study and

[4] The INCA website http://www.inca.org.uk/ is funded by the Qualifications and Curriculum Development Agency (QCDA) in England. Content is managed by the International Information Unit at the National Foundation for Educational Research (NFER).

attainment targets, GCSE qualification criteria,[5] and GCSE review of standards.[6]

5. Search for key words 'TIMSS', 'PIRLS', 'PISA' and 'validity' in assessment journals such as *Practical Assessment*, *Research and Evaluation*, *Assessment in Education*, *Psychological Bulletin*, *Research Notes*.

6. National statistics on key stage 2, key stage 3 and key stage 4 levels of achievement published online in the DCSF (2007) *National Statistics First Release*.

Despite the availability of primary data gathered by the IEA and the OECD, this research did not develop a method to carry out new statistical analysis. Yet, using the sources listed above, this study identified the salient themes in the literature on international assessing to generate hypotheses for further research. The available literature revealed patterns related to international assessing, helping the researchers to identify beliefs, and policies shaping the national and international assessments. In summary, the methodology used allowed the researchers to meet the challenge of contrasting national and international assessing.

Main argument

In October 2008 the government announced changes to the school accountability system and reform the assessment system at key stage 3. A small *ad hoc* Expert Group on Assessment (EGA) was established to advise on the proposed arrangements and their delivery. With a specialised task, they would give guidance on the next steps for key stage 3 assessment after the whole-cohort assessments were removed. They would also make recommendations on the development and delivery of a robust national sampling system to produce information about national standards in English, mathematics and science. This group was not a statutory body with a lengthy remit and was to operate strictly within the parameters set out by the secretary of state. The specialist group was to develop advice in the context of some fundamental principles. Assessing and assessment systems should give *parents* the information that they need in order to compare different schools, choose the right one

[5] QCA (2008) *GCSE qualification criteria*, QCA/07/3165, http://www.qca.org.uk/libraryAssets/media/qca-07-3469_gcse_qualification_criteria.pdf (accessed 9 February 2010).

[6] QCA (2007) *Review of Standards in GCSE English 2002-5.* February 2007, QCA/07/3102. http://www.ofqual.gov.uk/files/QCA-07-3102_standards_GCSE_English_mar07.pdf (acessed 23 February 2010)

for their child and then track their progress. Furthermore, England's assessment system should provide *teachers and heads* with the information that they need in order to assess the progress of every child and their school as a whole, without unnecessary bureaucracy. Finally, the system should allow the *public* to hold national, local government and governing bodies to account for the performance of schools.

In May 2009, the EGA published a report[7] that recommended that key stage 2 science assessments should be discontinued. In response to these recommendations, the government ended science national assessments for key stage 2 pupils. According to the EGA guidance, in May 2010 key stage 2 pupils would be taking their final high-stakes science assessments. From 2010, science at key stage 2 would be assessed through teacher assessment, and in view of such necessary changes, the EGA made a case for the integration of TIMSS, PIRLS and PISA into a more frequent cycle of national sample assessing. The argument was that although national assessments were no longer administered at key stage 3, and as science assessments would no longer be available for key stage 2 children, it was still important to ensure that the public and government would monitor average national performance at a crucial stage in pupils' learning. This monitoring would happen over an extended period of time through *national sample assessments*, which are radically different in nature compared to the national assessments sat by all children in England to account for their level of attainment in the national curriculum. According to the EGA, the new sample assessments should be taken at the end of year 9, and considering that England has participated in international comparison studies for many years and that PISA, PIRLS and TIMSS yielded valuable information on how England's pupils performed in comparison with those in other countries, 'where possible, assessment items should be linked to international comparison surveys in which England already participates (e.g. TIMSS)'.[8]

The EGA proposition to *link* the national sample assessments to international sample assessments was not as simple as it sounded because the types and the functions of national and international assessments had been arguably different for many years. Moreover, linked assessments ought to measure the same qualities. To be clear, a valid linkage between national and international assessments would require a common statistical specification behind the assessments to be linked. Hypothetical linkage would be valid only if common content specification was ensured at the assessment development stage. National and international assessments would need

[7] View the *Report of the Expert Group on Assessment* at http://publications.dcsf.gov.uk/eOrderingDownload/Expert-Group-Report.pdf (accessed 6 February 2010).

[8] The *Report of the Expert Group on Assessment,*, p. 9.

to sample similar content areas, comparable types of questions and 'similar ratios of questions to content areas sampled' (Newton, 2000, p. 10).

If the EGA proposition was to use statistical scaling to maintain national curriculum assessment standards from one year to the next, then international assessments would have to be adopted as anchor assessments. A great degree of rigour would be required during assessment construction. But unless the national assessments reproduced the international assessment specifications, using the international assessments as anchors would be impossible. National and international assessments would also have to be constructed with basis on a stable curriculum to protect the logic of scaling prior and subsequent assessments. Furthermore, valid links between national and international assessments would depend on technical criteria that include equal reliability of the assessment instruments and assessment administration under the same conditions (Feuer et al., 1998). As discussed in the following paragraphs, assessment researchers had been working on new strategies to maintain national curriculum standards. If some of the recommendations had been adopted, national and international assessments would perhaps measure the same construct, in similar ways.

From a historical perspective, the national curriculum assessments of England were originally conceived to be a census designed for political accountability purposes using statements of attainment on a 10-level scale to produce *levels of attainment* instead of marks. However, the national curriculum assessments seeking to measure levels of attainment were subsequently questioned by teachers and educationalists (Shorrocks-Taylor, 1999, p. 13) because they seemed unmanageable and held an excessive number of attainment targets. In the early 1990s there was also doubt about how performance on statements of attainment should be aggregated to represent performance on attainment targets. Statements of attainment were soon deemed as an unacceptable basis for standard setting (Newton, 2000). With the 1993 Dearing Review of the national curriculum and its assessment, the national curriculum assessments suffered major changes from being strongly 'criterion-referenced' based on clear *statements of attainment* to being weakly 'criterion-related' based on holistic *level descriptors* (Newton, 2000, pp. 1–6). Conversely, the international assessments were originally designed to be norm-referenced assessments and express the candidates' scores in rank order, based on a normal distribution curve. If the original model for the national curriculum assessments had not changed, and if the national curriculum assessments remained criterion-referenced assessments, then no one would ever have argued that they should be linked because there would be no statistical reason for it. The initial objective of the national curriculum assessments was to reveal the percentage of pupils who achieved the expected level in reading, writing, English, mathematics and science, including the percentage achieving the expected level in a combination of subjects. The national curriculum assessments were initially designed around set criteria to be

achieved. The standards applied referred to explicit performance criteria, whereby an individual candidate achieved criteria X, Y, Z. However, the mastery of learning aspect of the assessment is no longer the most important measurement aspect in the national curriculum assessments since the mid-1990s. Such changes to the national curriculum assessments could reinforce a case of linking national and international assessments, but the case for it would have been stronger if the national curriculum assessments were cohort- or norm-referenced, Also, one should bear in mind the historical background to PIRLS, TIMSS and PISA, and the 'limits of linking' examinations standards, as discussed in the literature.[9]

The international assessments always aimed to compare students' performance, informing us that certain students are better than others. The mean score and typical pattern of performance had been established by setting it to a large random sample of the age population. The performance by groups of pupils had been reported in terms of their position comparable to the mean and/or distribution of scores in the standardising sample. Yet it may be possible to request data on individual pupils' performance from assessment agencies if the need arises.

PIRLS, TIMSS and PISA are known to use similar linking methods between questions over time. The international assessments needed to be *linked* because they involve observation of some subset of a population of assessment items administered to groups of test takers. The sampling technique used required all assessment instruments to be linked so that ultimately performance of all pupils could be placed on a single scale using item response theory (IRT) methods.[10]

In summary, whereas international scores are interpreted with reference to the performance of the pupils taking the assessment, the national curriculum assessment results inform assessment administrators of pupils' levels of attainment in the national curriculum, but the national curriculum assessments are only weakly 'criterion-related'. Norm-referenced international assessments and weakly criterion-related national assessments are still grounded on different theoretical assumptions. However, when evaluating the main procedures for maintaining national curriculum assessment standards in key stage 2 English and science, Newton (2000, p. 2) once recommended alternative strategies that included a 'cohort reference national curriculum assessment results, not awarding levels at all, with the implication that school comparisons be based on the average standard score of pupils'. This would have approximated the national and international assessments.

---

[9] See Newton, P.E. (2005) 'Examination standards and the limits of linking', *Assessment in Education*, vol. 12, no. 2, pp. 105–23.

[10] NFER, *Background to PIRLS 2001*, http://www.teachernet.gov.uk/_doc/3980/PIRLS%20full%20report.pdf

Even if the national curriculum assessments still awarded levels whereas PIRLS, TIMSS and PISA awarded average standard scores of groups of pupils, national and international assessments could include similar types of questions, and perhaps key stage 2 assessment questions could be linked to international assessment questions. In this study we explain that there is a significant amount of curriculum overlap between national and international assessments, and assessment data may be related for the purpose of concurrent validation.

The international assessments are therefore definitely of interest to the EGA, and Ofqual now has strong reason to explain the relations between TIMSS, PIRLS and PISA in order to:

- address public concerns over the standards of assessments used as indicators of achievement in England (Standard 13.3)[11]
- keep the technical quality of the national assessments under check[12]
- complement national assessment results with information from other sources to generate defensible conclusions (Standard 15.4).[13]

Review of work in the area

A systematic comparison of the main characteristics of the key stage 2 and key stage 3 national assessments, GCSEs, TIMSS, PIRLS and PISA, has never been done before in the way that we do in this report. The reader will certainly find first-class studies commissioned by the DCSF to the NFER.[14] In the PIRLS 2001 *National*

[11] American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999, 2004) *Standards for educational and psychological testing*, p. 145.

[12] For a discussion on the theory of checks and balances, see the Introduction in Maravall, J.M. and Przeworski, A. (2003) *Democracy and the rule of law*, Cambridge University Press,. http://assets.cambridge.org/97805218/25597/sample/9780521825597ws.pdf (accessed 6 February 2010).

[13] American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999, 2004) *Standards for educational and psychological testing*, p. 167.

[14] Sturman, L., Ruddock, G., Burge, B., Styles, B., Lin, Y. and Vappula, H. (2008) *England's achievement in TIMSS 2007. National report for England*, Slough, NFER; DCSF Research Brief DCSF- RBX-18-08.

report for England, Twist et al.[15] wrote a detailed comparison of the PIRLS reading framework and the national curriculum. The authors found that the range of texts in the PIRLS assessments was narrower than that outlined in the national curriculum, mostly due to the requirements of translation. According to the same analysis, the PIRLS 2001 and 2006 assessments did not include the reading of genres usually present in the national curriculum assessment, such as playscripts, myth and newspaper articles. In the PIRLS 2006 *National report for England*, Twist et al.[16] mapped PIRLS processes of comprehension against the assessment focuses for reading and the reading skills being assessed in specific questions in national curriculum assessments in England. The PIRLS examination and evaluation of content, as well as language textual element, were not as detailed as the national curriculum assessment focuses. PIRLS did not ask children to explain and comment on writers' use of language, comment on grammatical and literary features, or identify and comment on writers' purposes and viewpoints. PIRLS did not ask children to relate texts to their social, cultural and historical contexts and literary traditions

Within the TIMSS 2007 *National report for England*, Sturman et al. (2008) noted how assessment items had been written to match an assessment framework with content domains. In the TIMSS grade 4 assessment, the content of papers was quite similar to the national curriculum in England, and the structure was also similar. TIMSS grade 8 included content domains of number, algebra, geometry, and data and chance like the national curriculum in England.[17] Considering assessment results, Sturman et al. (2008) concluded that England's national assessments results matched TIMSS 2007 outcomes, which proved to be 'broadly in line with gender findings from England's national assessments (GCSEs and key stage assessments taken in 2007)'.[18] Yet the GCSEs sat less comfortably with the PISA 2006 results in scientific literacy, which found rather more gender differences.

Likewise the national curriculum assessments, TIMSS, PIRLS and PISA have no personal consequences for test takers, whereas GCSEs have clear consequences on career prospects of test takers. However, it is undeniable that the national

[15] 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.

[16] Twist, L., Schagen, I. and Hodgson, C. (2007) *Readers and reading: the national report for England 2006* (PIRLS: Progress in International Reading Literacy Study), Slough, NFER, http://www.nfer.ac.uk/nfer/publications/PRN01/PRN01.pdf (accessed 6 February 2010).

[17] Sturman, L., Ruddock, G., Burge, B., Styles, B., Yin, L. and Vappula, H. (2008). *England's achievement in TIMSS 2007 (Trends in International Mathematics and Science Study)* (DCSF Research Brief RBX-19-8). London: DCSF

[18] Sturman, L., Ruddock, G., Burge, B., Styles, B., Yin, L. and Vappula, H. (2008), p. 95.

curriculum assessments impose political and administrative penalties on schools, so they involve greater preparation and anxiety if compared to the international assessments. Being low-stakes assessments, the international may be associated with a decrease in motivation and performance (Wise and DeMars, 2005; Wolf et al., 1995),[19] making those assessments hardly comparable. The issue of assessment-taking motivation could well challenge the validity of any comparison between student performance in the national and international assessments (Baumert and Demmrich, 2001).[20]

The Centre for Education and Employment Research (CEER) produced a report on PISA 2003, noting that the OECD had not done enough to demonstrate that its literacy assessments are measures of 'knowledge and skills for life', and that there was the suspicion that 'the maths and science assessments were more assessments of reading centred on elementary mathematical and scientific concepts'.[21] Such problems are not helpful for raising the profile of PISA assessments, making them akin to meticulous national assessment.

Additional technical problems with PISA (Smithers, 2004)[22]included:

- the relative standings of countries in TIMSS, PIRLS and PISA being definitely affected by the countries taking part, the types of questions asked, whether the target population was age-based or grade-based, and poor response rates in some countries.
- various sources of unintended bias, which include ignoring the degree of curriculum match, as TIMSS does
- the exclusion of difficult questions (by doing so, PISA caps high-performing educational systems)

[19] See Eklöf, H. (2007)*Test-taking motivation on low-stakes tests: a Swedish TIMSS 2003 example*, Department of Educational Measurement, Umeå University, Sweden; IERI Monograph Series, Issues and methodologies in large-scale assessments, http://www.ierinstitute.org/IERI_Monograph_Volume_01_Chapter_1.pdf (accessed 6 February 2010).

[20] Eklöf, H. (2007).

[21] See the Introduction to Smithers, A. (2004) 'England's education: what can be learned by comparing countries?', University of Liverpool, Centre for Education and Employment Research, http://www.suttontrust.com/reports/pisa_publication.doc (accessed 16 February 2010).

[22] Smithers, A. (2004), p.ii.

- that PISA inputs more meaning to the results than they actually hold – ranks of mean scores cannot be published as real ranks because they do not differ in significant ways

Moreover, an increase in school performance would not necessarily have a positive impact on the economy of industrialised countries (Robinson, 1999)[23] as PISA purports is the case.[24]

Furthermore, PISA has been blamed for a lack of concern with a notion of curriculum. Contents and skills assessed by PISA, said to be necessary later in life, are probably not the same as assessed by GCSEs, which are clearly curriculum-based. PISA, more than other international surveys, has been judged as an ambiguous assessment that assesses both end of schooling achievement and skills for the next stages of life. Therefore when the government of England favours PISA over TIMSS and PIRLS, because it can derive policy conclusions for its education system, it may regret this because PISA is already said to involve several factors, which incidentally include education.

Comparing TIMSS to PISA, Hutchison and Schagen (2006) note that international cross-sectional studies such as TIMSS and PISA have limited explanatory value for the government ministers and senior civil servants who use them to praise schools' attainment. The authors discuss some of the technical aspects of TIMSS and PISA, identifying discrepancies between TIMSS and PISA results at the country level. A major problem in both programmes was that they implicitly assumed a causal link between background factors without collecting longitudinal data to carry value-added analyses. Expected relationships between background factors and student performance cannot always be found in TIMSS and PISA surveys because 'correlation does not imply causation'. [25] One of the main problems with the three- to four-year cycle studies would be the lack of measure of recent progress. A possible solution for no measure of progress would depend on an agreement between governments and assessment agencies for the establishment of international longitudinal studies to produce data on progress over time.

---

[23] Robinson, K. (1999) *All our futures: creativity, culture and education*, National Advisory Committee on Creative and Cultural Education.

[24] Point made in *Education for the creative workforce: rethinking arts and education*, ARC Centre of Excellence for Creative Industries and Innovations in partnership with the Australia Council for Arts, March 2007, http://www.australiacouncil.gov.au/research/education_and_the_arts/reports_and_publications/creative_workforce_rethinking_arts_and_education .

[25] http://www.brookings.edu/gs/brown/irc2006conference/HutchisonSchagen_presentation.pdf (accessed 6 February 2010).

Goldstein (2008) and Bonnet (2002) reviewed some of the methodological issues surrounding international studies of comparative educational achievement, adding that studies from the OECD and the IEA should recognise cultural specificity in terms of assessment question development and in the subsequent analysis. Furthermore, the statistical models used in the analysis should incorporate multidimensionality to retain country differences rather than eliminate difference in favour of a common scale.

Comparisons

The following sections compare national and international assessment features, content specifications, measurement scales and curriculum match. As noted before, there is some reservation in the literature regarding comparisons of national and international assessments. A number of authors and statisticians managing high-stakes assessments may argue that national and international assessments measure different things under different conditions of preparedness. In order to assessment the robustness of such claims, the following sections seek evidence that the national and international assessments may sometimes position students on a same latent trait or on same underlying ability. Notwithstanding this, even if the national and international assessments do not match, they could still be used to validate each other, but first one needs to place the assessments and examinations side by side, in a systematic way.

Assessment features

This section compares the assessment features of PIRLS, TIMSS, PISA, national curriculum assessments and GCSEs regarding sample size, age of students, number of countries involved, cycles, first surveys conducted, and what these assessments measure. Whereas the national curriculum assessments were designed to measure standards of achievement in key areas of the national curriculum of England, TIMSS 2007 compared knowledge and understanding of mathematics and science. Likewise, PISA 2006 measured students' capacities to apply knowledge and skills in science, reading and mathematics literacy. Whereas PISA used a cohort referencing model to guarantee the same proportion of pupils achieving each of the levels or grades from one cycle to the next to serve the purpose of ranking countries' performance within a particular assessment administration, GCSEs measured candidates' ability to meet assessment objectives stated in the different GCSE subject criteria set out by Ofqual and followed by awarding bodies offering the qualification. [26]

[26] For a systematic and comprehensive source of information about developing test of knowledge, skills, and ability see Downing, S.M. and Haladyna, T.M. (2006)

PIRLS

The Progress in International Reading Literacy Study (PIRLS) is a cross-national study with waves, which informs governments as much as possible about students' strengths and weaknesses in reading comprehension. Assessment features could be summarised as in Table 1.

Table 1 PIRLS 2006 assessment properties

International sample size	215,155 pupils[27]
Age of students	9- to 10-year-olds, key stage 2 (4th grade, year 5)[28]
Number of countries	45 education systems (including 38 countries, 5 Canadian provinces, and the separate English- and French-speaking education systems in Belgium)
Cycle	Every 5 years
First survey	2001
What it measures	PIRLS measures children's *reading literacy* achievement by comparing the following processes of *reading comprehension*: (a) retrieving explicitly stated information; (b) making straightforward inferences; (c) interpreting and integrating ideas and information; (d) examining and evaluating content, language and textual elements.
Objective	To provide opportunities to make cross-national comparisons in terms of primary-age children's literacy and the factors associated with its acquisition, such as reading attitudes. Explain whether and what children read for pleasure. Enable countries to measure attainment relative to their own attainment five years earlier.
Participation and response rates in England	The achieved coverage of the nationally defined sample in England was 97.4 per cent (the international target was 95 per cent). A total of 4036 pupils in 148 schools in England were assessed in PIRLS 2006, with overall

[27] Source: Mullis, I.V.S., Martin, M.O., Kennedy, A.M. and Foy, P. (2007) *PIRLS 2006 international report*, Chestnut Hill, MA, TIMSS & PIRLS International Study Center, Boston College. See 'Appendix A: Supplementary information about PIRLS 2006 procedures', p. 294, http://pirls.bc.edu/PDF/P06_IR_AppendixA.pdf (accessed 6 February 2010).

[28] The sample participating in the survey is chosen based on *stage* rather than on *age*. Target population is defined as all students enrolled in the grade that represents four years of schooling, counting from the first year of ISCED (International Standard Classification of Education) level 1, provided that the mean age at the time of testing is at least 9.5 years.

	exclusions of 2.4% (1.6% school-level exclusions, and 0.9% within sample exclusions).[29]
Model of assessment (timing, available marks and question types)	The PIRLS 2006 reading material was divided into 40-minute 'blocks', each comprising a story or article and items representing at least 15 score points. There were eight such blocks, four for each reading purpose: literary and informational. The eight assessment blocks were distributed across ten assessment booklets, and each student completed one booklet in an 80-minute assessing session. Each booklet contained two blocks. Questions on the reading passages enabled pupils to demonstrate a range of abilities and skills in constructing meaning from written texts. PIRLS 2006 reported achievement results according to reading comprehension processes in addition to reading purposes.
Next survey	2011, with report being released in September 2012

The assessments and questionnaires used in the PIRLS 2006 study were developed by an international consortium and approved by all participating countries. The survey in England was conducted by the NFER and involved children in key stage 2, year 5.

By administering the assessment every five years, PIRLS allows countries to monitor their children's literacy achievement and collect background information about the students and schools. The reading achievement results present each country with an opportunity to examine educational policies and practices against a globally-defined benchmark.

PIRLS aims to provide a baseline for future studies of trends in reading literacy achievement, and to gather information about children's home and school experiences in learning to read. The PIRLS framework defines reading literacy as the 'ability to understand and use those written language forms required by society and/or valued by the individual' (Campbell et al., 2001, p. 3).[30]

[29] Source: IEA Progress in International Reading Literacy Study, Exhibit A.4 Coverage of PIRLS target population, in *PIRLS 2006 report*, Appendix A, p. 292.

[30] See reference to Campbell et al. (2001) at http://nces.ed.gov/pubs2004/pirlspub/2.asp?nav=1 (accessed 6 February 2010).

PIRLS passages and items generally have good psychometric characteristics, with a wide range of difficulty levels and good discrimination indices. Passages and assessment items are assigned to student assessment booklets according to a matrix sampling plan. The IEA ensures that the passages and items have good measurement properties in each country.

TIMSS

Like PIRLS, the Trends in International Mathematics and Science Study (TIMSS) is a cross-national, standardised international study of student achievement in mathematics and science organised by the IEA. Any country willing to take part and who can afford it can participate in such study. TIMSS targets students enrolled in the grades containing the largest proportion of 13-year-olds at the time of assessing. The TIMSS mission is to measure student achievement in mathematics and science in a way that does justice to the breadth and richness of these subjects as they are taught in the participating countries. TIMSS monitors countries' improvement or decline by tracking trends in student performance from one assessment cycle to the next. Planning is underway for a new survey in 2011. The assessment properties can be summarised as in Table 2.

Table 2 TIMSS 2007 assessment properties

Sample size internationally	183,136 pupils (4th grade, year 5) 241,613 pupils (8th grade, year 9)[31] Total: 424,749
Age of students	Population 1: 9- to 10-year-olds, 4th grade (year 5) Population 2: 13- to 14-year-olds, 8th grade (year 9)
Number of countries	59 countries
Cycle	Every 4 years
First survey	1995
What it measures (at 4th and 8th grades)	TIMSS measures achievement at the 4th and 8th grades in different content domains and cognitive domains. TIMSS also monitors curricular implementation in the participating countries, identifying promising instructional practices from around the world. TIMSS mathematics 4th grade measures achievement in three content domains (number, geometric shapes, data display) and three cognitive domains (knowing, applying, reasoning). TIMSS mathematics 8th grade measures achievement in four content domains (number, algebra, geometry, data and chance) as well as measuring the same three cognitive domains (knowing, applying, reasoning). TIMSS science 4th grade measures achievement in three content domains (life science, physical science, earth science) and the same cognitive domains as above. TIMSS science 8th grade measures achievement in four content domains (biology, chemistry, physics, earth

[31] Source: Mullis, I.V.S., Martin, M.O. and Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A. and Galia, J.) (2008a) *TIMSS 2007 International mathematics report: findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*, Chestnut Hill, MA, TIMSS & PIRLS International Study Center, Boston College, pp. 390–1, http://pirls.bc.edu/TIMSS2007/PDF/T07_M_IR_AppendixA.pdf (accessed 6 February 2010).

	science) and the same cognitive domains as above.
Objective	To provide opportunities to make cross-national comparisons in terms of: (a) students' knowledge and understanding of mathematics and science; (b) students' and teachers' attitudes towards these subjects; (c) mathematics curriculum; (d) science curriculum; (e) teaching conditions and practice.
Participation and response rates in England	In England, 143 schools and 4316 year 5 pupils (4th grade) and 137 schools and 4025 year 9 pupils (8th grade) participated in TIMSS 2007. These numbers represented 98 percent of the year 5 sample and 88 percent of the year 9 sample of pupils, as well as 90 percent of participating schools (year 5) and 86 percent% of participating schools (year 9).[32]
Model of assessment (timing, available marks and question types)	TIMSS 2007 materials consisted of 28 blocks of items, distributed across 14 student booklets. Each booklet consisted of four blocks of items. Students answered assigned booklets within 72 minutes at 4th grade and 90 minutes at 8th grade. TIMSS 2007 assessment time was comparable to that in the 1995, 1999 and 2003 assessments. To enable linking between booklets, each block of questions appeared in two booklets. The booklets were organised into two two-block sessions (Parts I and II), with a break between the parts. The content of the written assessments was rotated to provide eight different versions, providing wide curriculum coverage. Each assessment, numbered 1–8, contained a mixture of both science and mathematics questions. In England, the written assessments were presented to students in two separate booklets, so each student completed an A and B booklet, including mathematics and science questions arranged in clusters. The student's questionnaire formed a third booklet.

[32] Source: NFER, http://www.nfer.ac.uk/research/projects/trends-in-international-mathematics-and-science-study-timss/

Next survey	2011, with international reporting for both year levels scheduled for December 2012

TIMSS 2007 was the fourth survey in a linked series, providing information on trends in performance over time. The 2007 survey involved 143 primary schools and 137 secondary schools in England. These studies produced a robust picture of performance for pupils in key stage 2 or 4th grade (England's year 5) and key stage 3 or 8th grade (year 9).

PISA

The Programme for International Student Assessment (PISA) is a cross-national, cross-sectional, internationally standardised study in which only OECD countries plus 'partner' countries are accepted to participate; thus it is a selective study if compared to TIMSS and PIRLS. PISA targets a very specific population of 15-year-old students attending educational institutions located within the country. Such students can be in grade 7 or higher, but they must be enrolled in an educational institution. PISA is developed by participating countries and administered in three-year cycles. Apart from assessing mathematics and science literacy to investigate what students from diverse countries can do with what they have learned at school, PISA also surveys the attitudes and opinions of students in relation to learning. The PISA survey was implemented in 43 countries in the first wave in 2000, in 41 countries in the second wave in 2003, and in 57 countries in the third wave in 2006, and 62 countries have signed up to participate in the fourth survey in 2009. PISA assessments are typically administered to between 4500 and 10,000 students in each country.

PISA 2006 was a response to the need for cross-nationally comparable evidence on student performance, and it was defined by the OECD as 'a commitment by governments to monitor the outcomes of education systems in terms of student achievement on a regular basis and within an internationally agreed common framework'.[33] The OECD published data on 15-year-olds' achievements only for the UK. However, separate results for England, Wales, Northern Ireland and Scotland are accessible in the NFER PISA 2006 national report.[34]

[33] PISA (2006) *Science competencies for tomorrow's world*, Vol. 1, *Analysis*, OECD, p. 3.

[34]See Bradshaw, J., Sturman, L., Vappula, H., Ager, R. and Wheater, R. (2007) *Achievement of 15-year-olds in England: PISA 2006 national report* (OECD Programme for International Student Assessment), Slough, NFER, http://www.nfer.ac.uk/nfer/publications/NPC02/NPC02.pdf

Table 3 PISA 2006 assessment properties

Sample size internationally	400,000 students, representing almost 20 million 15-year-olds enrolled in the schools of the participating countries in 2006
Age of students	15-year-olds as they approach the end of compulsory education, getting ready to take their GCSEs in England
Number of countries	57 (30 OECD member countries and 27 partner countries)
Cycle	Every 3 years
First survey	2000
What it measures	PISA measures students' capacities to apply knowledge and skills in science, reading and mathematical literacy. PISA science, for example, measures the following knowledge domains: physical systems, living systems, earth and space systems, technology systems, scientific enquiry and scientific explanations. Competencies involved are: identifying scientific issues, explaining scientific phenomena and using scientific evidence. PISA also assessments the areas of application of science in relation to health, natural resources, environment, hazard, frontiers of science and technology. In 2006 more assessment time was given to scientific literacy than to mathematics and reading literacy. The sequence of main focus subjects of PISA so far is 'reading' (2000), 'mathematics' (2003), 'science' (2006) and 'reading' (2009).
Objective	PISA aims to supply regular information on educational outcomes within and across countries by providing insight about the range of skills and competencies in different assessment domains of literacy, which are considered to be essential to an individual's ability to participate and contribute to society. Such domains are broken down in terms of (1) the content or structure of knowledge that students need to acquire, (2) the processes that need to be performed, (3) the contexts

	in which knowledge and skills are applied. PISA seeks to inform policy makers on what educational structures and practices maximise the opportunities of students from disadvantaged backgrounds. It also looks at the extent to which student performance is dependent on background.
Participation and response rates in England	In England, 169 schools and 4935 students participated in PISA 2006, which represented 89 percent of sampled schools and 89 per cent of sampled students. The weighted school response for the combined UK sample was 88 per cent. This was just 1 per cent below the target participation rate. This was a great improvement on previous PISA surveys in the UK.[35]
Model of assessment (timing, available marks and question types)	Pencil-and-paper assessments, lasting 2 hours for each student. PISA 2006 used both multiple-choice items and questions requiring students to construct their own answers. Stimulus materials included texts, tables and graphs, followed by questions on various aspects of the text, table or graph, with the questions constructed so that the tasks were similar to those encountered in the real world. Items were typically organised in units based on a passage describing a real-life situation. A total of 6.5 hours of assessment items was included, with different students taking different combinations of the assessment items. Assessment booklets contained questions from one or more of the scientific, reading and mathematical literacy domains. PISA 2006 included 210 minutes of science material. Students also answered a 30-minute questionnaire about their background.
Next survey	NFER conducted the PISA 2009 survey in England, Wales and Northern Ireland in November 2009, with results scheduled for release at the end of 2010. PISA 2012 is included in the government agenda.

[35] Bradshaw et al. (2007, p. vii).

In the OECD's perspective, some of the key features driving the development of PISA are:

- policy orientation – PISA identifies the characteristics of schools and education systems that have high performance standards, establishing benchmarks for educational improvement
- regularity – PISA assessments allow for countries to monitor their key learning objectives
- wide geographical coverage of a collaborative programme – PISA has become a major assessment tool beyond the OECD countries.

Candidates from the UK who participate in the PISA survey are approximately in the same age group as those who sit GCSEs.

National curriculum assessments

National curriculum assessments (NC assessments) are longitudinal census, which assess pupils' attainment in English and mathematics at key stages 1, 2 and 3, and science at key stages 2 and 3, although from 2009 the key stage 3 NC assessments were no longer statutory.

NC assessments have both standardised syllabuses and prescribed criteria, and their results are intended to complement evidence of attainment collected by teachers via their own assessments. The levels of performance in the NC assessments are attributed by matching pupils to *level descriptors* by means of a process of 'best fit' to inform parents and teachers of individual pupil progress. When aggregated, the NC assessment results indicate the level of performance of schools and local authorities in England.

Table 4 The national curriculum assessment properties[36]

Cohort	Key stage 2 English – 650,000 pupils (2007) Key stage 2 mathematics – 650,000 pupils (2007) Key stage 2 science – 650,000 (2007)[37] In 2007 the NC assessments included all pupils attending maintained schools in England who were working within the levels of achievement specified. All pupils had to undertake the national assessments in English, mathematics and science at the end of key stages 2 and 3 (at ages 11 and 14). In 2007 this amounted to well over 90 of the cohort. Every year over 600,000 pupils take the national assessments.
Age of students	Key stage 2 assessments target year group 6: 11-year-olds Key stage 3 assessments target year group 9: 14-year-olds
Number of countries	One
Cycle	Yearly (May)
First survey	1995
What it measures	Pupils' attainment is assessed in relation to the national curriculum programmes of study,[38] including English, mathematics and science.[39] Pupils achieving government-set attainment targets[40] are awarded levels

[36] Years of reference 2007 and 2008.

[37] QCA Test Statistics 2007, http://assessmentsandexams.qca.org.uk/libraryAssets/media/2007_KS2_test_statistics.pdf (accessed 6 February 2010).

[38] Programmes of study can be found at http://curriculum.qca.org.uk (accessed 6 February 2010).

[39] You can view the English and mathematics key stage 2 programmes of study at http://curriculum.qca.org.uk/key-stages-1-and-2/subjects/index.aspx (accessed 6 February 2010), and the English, mathematics and science key stage 3 programmes of study at http://curriculum.qca.org.uk/key-stages-3-and-4/index.aspx (accessed 6 February 2010).

[40] Key stage 2 attainment targets for key stage 2 English can be found at http://curriculum.qca.org.uk/key-stages-1-and-

	on the national curriculum scale to reflect their attainment. NC assessments are a measurement of achievement against the precise attainment targets of the national curriculum rather than any generalised concept of ability in any of the subject areas.
Objective	NC assessments are designed to measure standards of attainment reached by children in key areas of the curriculum.
Participation	In 2007 the NC assessments covered all schools in England with pupils eligible for assessment at key stage 2 and key stage 3[41] in 2008, but as participation by independent schools is voluntary, the national analyses include only results from those independent schools that chose to make a return. Ofqual sets the Code of Practice that governs all aspects of the assessment process, from the development of assessments to the collection and reporting of results data. Ofqual monitors the system, which must comply with criteria in the Code of Practice.
Model of assessment (timing, available marks and question types)	Pen-and-paper assessments. Key stage 2 English: Two assessments: (1) reading (45 minutes, plus 15 minutes of reading time) [50 marks]; (2) writing [50 marks], composed of a longer writing task (45 minutes), a shorter writing task (20 minutes) and spelling (10 minutes). Key stage 3 English: Three papers assessing aspects of reading, writing, spelling and handwriting: (1) reading (15 minutes reading plus 1 hour to answer questions);]; (2) writing, comprising: longer writing task and shorter writing task (including spelling) (1 hour and 15 minutes); (3) Shakespeare paper (reading) (45 minutes). Marks from all three papers are aggregated to calculate the overall *English* Level

2/subjects/english/attainmenttargets/index.aspx?return=/key-stages-1-and-2/subjects/english/keystage1/index.aspx%3Freturn%3D/key-stages-1-and-2/subjects/index.aspx (accessed 6 February 2010).

[41] The key stage 3 NC tests in English, mathematics and science are no longer statutory. The tests were freely available for all maintained and independent schools for 2009 if required.

	The key stage 2 English assessment is keyed to assessment focuses based on the level descriptors. There are tier arrangements for key stage 3 written papers: tier 3–5, tier 4–6, tier 5–7, and tier 6–8.
Next survey	Key stage 2 implemented in May 2010, etc. Key stage 3 assessments no longer statutory from 2009.[42]

The national curriculum assessments are developed by specialist agencies with the help of subject experts and former teachers.

GCSEs

The Education Reform Act 1988 introduced the General Certificate of Secondary Education (GCSE) as the main school-leaving qualification taken by 14- to 16-year-olds in England, Wales and Northern Ireland. More than 700 qualifications are available on several subjects to any centre willing to offer them. The knowledge, skills and understanding that pupils acquire at key stage 3 make up the basis for future learning through GCSE specifications. The GCSE specifications offer schools a wide choice of content, but all must conform to the requirements set out in the GCSE and A level *subject criteria*,[43] which include common assessment objectives. These are designed to build on the key concepts and key processes set out in the key stage 3 programmes of study. Ofqual, the Department for Children, Education, Lifelong Learning and Skills (DCELLS) and the Council for the Curriculum Examinations and Assessment (CCEA) are responsible for the qualification criteria and subject criteria for GCSEs. Such qualification criteria set the basic rules on the structure of GCSEs and their assessment and grading.

[42] On 14 October 2008, the Secretary of State announced that the key stage 3 national curriculum assessments were no longer statutory for 2009. The key stage 3 national curriculum assessments would be replaced by improved classroom assessment by teachers and frequent reporting to parents in years 7, 8 and 9. View http://testsandexams.qcda.gov.uk/libraryAssets/media/Changes_to_national_curriculum_tests_and_teacher_assessments.pdf (accessed 6 February 2010).

[43] See GCSE subject criteria at http://www.ofqual.gov.uk/743.aspx (accessed 6 February 2010).

Subject criteria provide the framework within which awarding organisations[44] create the detail of GCSE specifications. The criteria set out the knowledge, understanding, skills, assessment objectives and scheme of assessment common to all specifications in a given subject. For subjects offered by more than one awarding body, the criteria ensure that there is comparability between specifications.

Table 5 GCSE assessment properties 2006–8

Cohort	721,577 (12.76%) sitting English in 2008 738.451 (13.02%) sitting mathematics in 2008 558.387 (9.85%) sitting English Literature in 2008[45] In total, 656,667 15-year-olds sitting GCSE English and mathematics in 2006/07[46]
Age of students	14- to 16-year-olds
Number of countries	England, Wales and Northern Ireland
Cycle	Yearly exams usually take place in January and May/June, with the majority in June
First survey	1988[47]
What they measure	GCSEs measure candidates' ability to meet assessment objectives (AOs) stated in each subject specification. AOs are stated in the different GCSE subject criteria (http://www.ofqual.gov.uk/743.aspx) whereby Ofqual and QCDA set out the knowledge, understanding, skills and assessment objectives common to all GCSE specifications in different subject areas, providing the framework within

[44] There are five exam boards that offer GCSE qualifications: AQA (www.aqa.org.uk), CCEA (www.ccea.org.uk), Edexcel (www.edexcel.org.uk), OCR (www.ocr.org.uk) and WJEC (www.wjec.co.uk) (all these websites accessed 6 February 2010).

45 Source: Joint Council for Qualifications (JCQ) Entry Trends 2008 – GCSE, Applied GCSE and Entry Level http://www.jcq.org.uk/attachments/published/1022/GCSE~AppGCSE~Entry%20Trends.pdf

46 Source:. . Joint Council for Qualifications (JCQ) Entry Trends 2008 – GCSE, Applied GCSE and Entry Level.

[47] In 1988, GCE O levels were phased out in favour of the General Certificate of Secondary Education (GCSE). See Robinson, C. (2007) 'Awarding examination grades: current processes and their evolution', in Newton, P., Baird, J., Goldenstein, H., Patrick, H. and Tymms, P. (2007) *Techniques for monitoring the comparability of examination standards*, London, QCA, p. 110.

	which an awarding body creates the detail of the specification. Students who complete GCSEs at level 1 can move on to other courses or work-based training at levels 1 or 2. A number of key skills may be assessed through the GCSE course content and the related scheme of assessment, as defined in each specification. Candidates may demonstrate their ability to fulfil aspects of the following key skills, at the level at which the qualification is accredited (level 1, 2 or 3).[48] Candidates must demonstrate abilities in (1) communication, (2) information technology, (3) improving on learning and performance, (4) working with others, and (5) problem-solving. Grade descriptors are provided by the regulator to give a general indication of the standards of achievement likely to have been shown by candidates awarded particular grades. GCSEs are graded A*–G and U (unclassified): higher tier exams lead to grades A*–D, and foundation tier exams lead to grades C–G. GCSEs were being revised in 2008, and from 2010 they will thoroughly assess functional skills.
Objective	GCSEs are qualifications offered by exam boards – organisations or consortia recognised by the regulatory authority for the purpose of awarding specified qualifications. Students aiming at higher education may need GCSEs in certain subjects. Most universities and colleges demand five GCSEs at grades A*–C, including English and mathematics (as well as A levels or equivalent qualifications). Preparation for the qualification mainly involves studying the theory of a subject, combined with some investigative work. Some subjects also involve practical work. GCSEs are usually studied full-time at school or college. All learners must achieve an average of 120 guided learning hours (GLH) or

[48] Different types of qualifications are grouped together into various 'levels'. All qualifications assigned the same level are broadly comparable with each other. GCSE levels indicate that one type of qualification can lead on to another. Yet a qualification may be made up of units that are not all regarded as being at the same level.

	more of additional and specialist. The GLH for GCSEs usually range from 90 to 145 hours.[49]
Participation	Up to 4.192.856 14 to 16 year olds (entries in 2008). Each year around 6.5 million GCSE qualifications are awarded.
Model of assessment (timing, available marks and question types)	GCSEs are assessed mainly on written exams, although in some subjects there are also elements of coursework. Some subjects, like art and design, have more coursework and fewer exams. Some GCSE courses are made up of units with different weightings; for these, students take exams at the end of each unit. Other GCSEs involve exams at the end of course. For some subjects, everyone sits the same exam. For others, students have a choice of two tiers: 'higher' or 'foundation'. Each tier leads to a different range of grades. Subject teachers normally decide which tier is best for each student. Examiners work out how many 'raw marks' test takers need to get a certain grade. If a student has taken a GCSE made up of units, the results slip may show a point scores on the uniform mark scale (UMS). The UMS is a system that examiners use to combine different unit marks to determine the overall GCSE grade. Because the question papers are different on every occasion, the marks do not retain common standards.

GCSEs went through changes in 2007 when QCA, in collaboration with teachers, awarding bodies, subject associations, higher education organisations and other interested parties, revised subject criteria to update the content of GCSEs and encourage innovative teaching, learning and assessment. By incorporating key elements of 14–19 curriculum developments, QCA ensured that the revised GCSEs complemented the new Diplomas. Moreover, assessment arrangements were revised to provide stretch and challenge for all learners, thus maintaining standards.

[49] Guided learning hours (GLH) are the number of hours of teacher-supervised or directed study time required to teach a qualification. The guided learning hours for GCSEs usually range from 90 to 145 hours, although the current normal value is set at 120 hours, which represents the amount of direct teaching support required by a typical learner to complete a GCSE programme, not the amount of learning time that a learner would require.

As part of the reform of 14–19 education, GCSE qualifications were again being reviewed by the regulators and other interested parties in 2009. There would be changes to the way GCSEs were assessed from September 2009. Coursework in most subjects would be replaced by controlled assessments, supervised by teachers in school. From September 2010, GCSE English, mathematics, and information and communications technology (ICT) would place more emphasis on the essential skills for work and adult life.

Content specifications

There are fundamental differences between PIRLS, TIMSS and PISA in relation to national curriculum assessments and GCSEs insofar as the international assessments are cross-sectional studies and the national assessments are longitudinal studies. Yet both are observational studies and the differences between national and international studies do not have a necessary impact on content specifications. On one hand, cross-sectional studies such as TIMSS, PIRLS and PISA must be based on a representative sample of test takers drawn from a population of students at one point in time. On the other hand, longitudinal studies such as the national curriculum assessments and the GCSEs involve a series of measurements taken over a period of time. Unlike the international assessments, the national curriculum assessments and GCSEs follow the experiences and outcomes of education overtime of a cohort of students defined with basis on shared experience. That is, the international assessments provide only a 'snapshot' of the performance of national samples of pupils at a particular point in time taking a subset of a population of items. Moreover, PIRLS, TIMSS and PISA tend to compare groups of test takers at different ages with respect of independent variables, such as course content that is actually taught. The international surveys also investigate the influence of textbooks and the training of teachers, whereas the national curriculum assessments evaluate the performance of all pupils based on holistic judgments of pupils' abilities at different levels as described in the national curriculum of England.

Despite the cross-sectional nature of their international surveys, IEA and OECD guarantee a meaningful connection between different waves of assessments designed to measure trends in educational achievement over time. Assessment items and content specifications are linked, and reproduced, to provide information on same age pupils' achievement on the same questions over time. PIRLS, TIMSS and PISA are known to use similar linking methods between questions over time.[50] For example, PIRLS 2006 contained ten blocks of items of which four were common

[50] See OECD - *PISA 2006 Technical Report* at http://www.pisa.oecd.org/dataoecd/0/47/42025182.pdf See also IEA -*PIRLS 2011 Assessment Framework* http://timss.bc.edu/timss2011/downloads/TIMSS2011_Frameworks-Chapter4.pdf and IEA - *TIMSS 2011 Assessment Design* at http://timss.bc.edu/pirls2011/downloads/PIRLS2011_Framework.pdf

to the 2001 and 2006 assessments, so that any changes could be based on linking these item blocks. TIMSS 2007 also used a matrix-sampling approach that involved bunching the entire assessment pool of mathematics and science questions into 14 student achievement booklets.[51] According to Sturman et al. (2008) the actual content of TIMSS 2007 was quite similar to the national curriculum in England. The similarity could be shown by looking at the topic 'number'.[52]

This section compares the content specifications of the national and international assessments, explaining why the content specifications of the national curriculum assessments and GCSEs may overlap but the content specifications of the international assessments tend to be narrower.

PIRLS

The PIRLS Reading Development Group (RDG) and National Research Coordinators (NRCs) from countries involved in PIRLS developed PIRLS 2006 reading assessments with a focus on three main areas of literacy: process of comprehension, purposes for reading, and reading behaviours and attitudes. In addition, PIRLS used a background questionnaire. The written assessment was designed to address the process of comprehension via the two purposes for reading: reading for literary experience and reading to acquire and use information. Being a norm-referenced assessment, PIRLS content was selected according to how well it would rank pupils from high to low achievers.

In 2006 the PIRLS assessment devoted 50 per cent of the questions to each reading purpose and process – 'literary experience' and the 'acquisition and use of information' (Table 6).

[51] Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., Arora, A. and Erberber, E. (2008) *TIMSS 2007 assessment frameworks*, p. 97, http://timss.bc.edu/TIMSS2007/PDF/T07_AF.pdf (accessed 19 February 2010).

[52] Sturman, L., Ruddock, G., Burge, et al. (2008) *England's achievement in TIMSS 2007: national report for England*, Slough, NFER, Chapter 5, p. 65, http://www.nfer.co.uk/nfer/publications/TMO01/TMO01.pdf (consulted 19 February 2010)

Table 6 Reading purposes and processes, PIRLS 2006

PIRLS 2006 – reading purposes and processes of comprehension	Literary experience	Acquire and use information	
Total	50%	50 %	100%
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%

Source: Twist, L., Schagen, I. and Hodgson, C. (2007, p. 118), http://www.nfer.ac.uk/nfer/publications/PRN01/PRN01.pdf (accessed 8 February 2010).

The literary and information texts used in PIRLS 2006 were all full-length stories or information pieces of 400–700 words. Questions covered the four processes of comprehension listed in Table 6. To capture pupils' responses, PIRLS included multiple choice questions and open response format, in normal ways used by pupils in England.

The choice of literary texts was not always challenging for English pupils.[53] The 'Culturally Balanced Assessment of Reading' (C-Bar), a European project commissioned by the European Network of Policy Makers for the Evaluation of Education Systems, is a key reference for those interested on debates regarding the validity of literary texts as presently used for all test takers, independently of their cultural backgrounds.[54]

[53] Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O., and Sainsbury, M. (2001). *Framework and Specifications for PIRLS Assessment 2001.* Chestnut Hill, MA: Boston College. Chapter 2, pp.9-20. http://timssandpirls.bc.edu/pirls2001i/pdf/PIRLS_frame2.pdf

Sainsbury, M. (2001) *NFER PIRLS 2001 report*, Chapter 4 'The PIRLS reading literacy tests', http://www.teachernet.gov.uk/_doc/3980/PIRLS%20full%20report.pdf (accessed 8 February 2010).

[54] European Network of Policy Makers for the Evaluation of Education Systems, http://cisad.adc.education.fr/reva/pdf/cbarfinalreport.pdf.

It should be noted that the PIRLS assessments give uneven weights to different reading processes. For example, in PIRLS 2006, the interpretation and integration of ideas and information was worth 61 marks, whereas examination and evaluation of content, language and textual elements was worth only 23 marks (Table 7).

Table 7 Distribution of PIRLS 2006 items by reading process

PIRLS 2006 – items by reading process	Total number of items	Number of multiple choice items	Number of constructed response items	Marks
Focus on and retrieve explicitly stated information	31	19	12	36
Make straightforward inferences	43	29	14	47
Interpret and integrate ideas and information	34	6	28	61
Examine and evaluate content, language and textual elements	18	10	8	23
Total	126	64	62	167

TIMSS

TIMSS 2007 measured two dimensions of performance for mathematics and science. Assessment contents were chosen on the basis of how well they discriminated among pupils. The TIMSS framework specified the subject matter and the thinking process to be assessed[55] at each grade, as outlined in Table 8.

[55] For a thorough explanation of the cognitive domains, see Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., and Brenwald, S. (2008). *Highlights From TIMSS 2007Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2009–001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington,DC., Table 5, p. 10, http://nces.ed.gov/pubs2009/2009001.pdf (accessed 19 February 2010).

Table 8 Content and cognitive domains by grade level, TIMSS 2007

4th grade content domains Mathematics	Percentages Mathematics	Percentages Mathematics	4th grade content domains Science	Percentages Science
Number	50%	–	Life science	45%
Geometric shapes and measures	35%	–	Physical science	35%
Data display	15%	–	Earth science	20%
8th grade content domains Mathematics	Percentages Mathematics	Percentages Mathematics	8th grade content domains Science	Percentages Science
Number	30%	–	Biology	35%
Algebra	30%	–	Chemistry	20%
Geometry	20%	–	Physics	25%
Data and chance	20%	–	Earth science	20%
Cognitive domains	4th grade	8th grade	4th grade	8th grade
Knowing	40%	35%	40%	30%
Applying	40%	40%	35%	35%
Reasoning	20%	25%	25%	35%

Source: http://timss.bc.edu/TIMSS2007/PDF/T07_TR_Chapter2.pdf (accessed 8 February 2010).

Mathematics

TIMSS 2007 mathematics 4th grade assigned greater emphasis to 'Number'. At 4th grade, 50 per cent of the assessment was on 'Number', whereas at 8th grade only 30 per cent covered the same content. The cognitive domains assessed by TIMSS 2007 mathematics were the same for both grades, encompassing a range of processes involved in working mathematically and solving problems through the primary- and middle-school years.[56] In terms of the percentage of time in mathematics class devoted to TIMSS content domains during the 2007 school year, year 5 pupils at school in England had attended more lessons on 'Number' (56 per cent) than the international average (50 per cent), but had spent less time on 'Geometric shapes and measures' (22 per cent) than the international average (24 per cent). The percentage of time spent on TIMSS 2007 topics was as set out in Table 9.

Table 9 Time (%) in mathematics class devoted to TIMSS content domains, 2007 (4th grade)

Country	Number	Geometric shapes and measures	Data display	Other	Total
England KS2, 4th grade	56	22	18	4	100%
International average 4th grade	50	24	16	10	100%

Source: http://timss.bc.edu/TIMSS2007/PDF/T07_M_IR_Chapter5.pdf (accessed 8 February 2010), p. 196.

The distribution of instructional mathematics time at 8th grade (year 9) in England was slightly different from the international average and also different from Singapore, one of the best performing countries in TIMSS 2007. Yet instructional time at 8th grade (year 9) was higher in England than the average of international teaching on topics such as 'Number', but lower than the average in 'Algebra' (Table 10).

[56] *TIMSS 2007 mathematics framework*, pp. 374–5, http://timss.bc.edu/TIMSS2007/PDF/T07_M_IR_AppendixA.pdf (accessed 8 February 2010).

Table 10 Time (%) in mathematics class devoted to TIMSS content domains, 2007 (8th grade)

Country	Number	Algebra	Geometry	Data and chance	Other	Total
England 8th grade	28 (0.7)	27 (0.6)	21 (0.4)	20 (0.4)	4 (0.5)	100%
International average 8th grade	24 (0.1)	29 (0.1)	27 (0.1)	13 (0.1)	7 (0.1)	100%

Source: http://timss.bc.edu/TIMSS2007/PDF/T07_M_IR_Chapter5.pdf (accessed 8 February 2010), p. 197.

Science

At both 4th and 8th grades, the TIMSS 2007 science assessment was also organised around two dimensions: a content dimension and a cognitive dimension. At 4th grade, the three subject areas to be assessed were 'life science', 'physical science' and 'earth science' (with 45 per cent of the assessment on 'life science'). At 8th grade, the four subject areas to be assessed were 'biology', 'chemistry', 'physics' and 'earth science', with more emphasis (35 per cent) given to 'biology'.

The TIMSS mathematics and science frameworks published by IEA described the topic areas covered and the objectives within each topic, as partially demonstrated in Table 11. Looking at the TIMSS 2007 8th grade mathematics framework, the subject matter 'Numbers' was less prevalent than in the 4th grade mathematics framework.

Table 11 Content domains, mathematics, TIMSS 2007

4th grade (year 5)	8th grade (year 9)
Number: whole numbers 1. Represent whole numbers using words, illustrations or symbols. 2. Demonstrate knowledge of place value, including recognising and writing numbers in expanded form. 3. Compare and order whole numbers. 4. Know the four operations (+, −, ×, ÷) and compute with whole numbers. 5. Recognise multiples and factors of numbers; read weight and temperature scales marked in multiples. 6. Estimate computations by approximating the numbers involved. 7. Solve problems, including those set in real-life contexts. 8. Solve problems involving proportions.	**Number: whole numbers** 1. Demonstrate knowledge of place value and of the four operations. 2. Find and use multiples and factors of numbers, read scales, and identify prime numbers. 3. Use principles of commutation, association and distribution. 4. Evaluate powers of numbers and square roots of perfect squares to 144. 5. Solve problems by computing, estimating or approximating.

Number: fractions and decimals	**Number: fractions and decimals**
1. Recognise fractions as parts of unit wholes, parts of a collection, locations on number lines, and divisions of whole numbers. 2. Represent fractions using words, numbers or models. 3. Identify equivalent fractions; compare and order fractions. 4. Add and subtract simple fractions. 5. Show understanding of decimal place value, including recognising and writing decimals using words and numbers. 6. Add and subtract decimals. 7. Solve problems involving simple fractions or decimals.	1. Compare and order fractions and decimals. 2. Demonstrate knowledge of place value for decimals. 3. Represent decimals and fractions, and operations with decimals and fractions, using models (eg number lines); identify and use such representations. 4. Recognise and write equivalent fractions. 5. Convert between fractions and decimals. 6. Compute with fractions and decimals. 7. Solve problems by computing, estimating and approximating.

Source: TIMSS 2007 report, mathematics framework, http://timss.bc.edu/TIMSS2007/PDF/T07_AF.pdf (accessed 8 February 2010).

'Scientific inquiry' was treated as an overarching assessment strand in TIMSS 2007. It overlapped all of the fields of science and had both content- and skills-based components. Assessment of scientific inquiry included items and tasks requiring students to demonstrate knowledge of the tools, methods and procedures necessary to (1) do science, (2) apply this knowledge to engage in scientific investigations, and (3) use scientific understanding to propose explanations based on evidence. The science framework expected students at both grade levels to possess some general knowledge of the nature of science and scientific inquiry. In addition to general knowledge, students should demonstrate the skills and abilities involved in five major aspects of the scientific inquiry process.

The TIMSS 2007 science framework provided an overview of the topic areas to be covered in each domain, followed by a set of assessment objectives for each topic area.[57] These objectives were written in terms of behaviours to be elicited (Table 12).

Table 12 Content domains, science, TIMSS 2007

4th grade (year 5)	8th grade (year 9)
Life science 1. Characteristics and life processes of living things 2. Life cycles, reproduction, and heredity 3. Interaction with the environment 4. Ecosystems 5. Human health *Examples of key topics:* 1. Distinguish between living and non-living things; identify common features of living things (movement; basic needs for air, food, water; reproduction; growth; response to stimuli). 2. Compare and contrast physical and behavioural characteristics of major groups of organisms (eg insects, birds, mammals, plants), and identify or provide examples of plants and animals belonging to these groups. 3. Relate major body structures in humans and other organisms (plants and animals) to their functions (eg digestion takes place in the stomach, teeth break down food, bones support the body,	**Biology** 1. Characteristics, classification and life processes of organisms 2. Cells and their functions 3. Life cycles, reproduction and heredity 4. Diversity, adaptation and natural selection 5. Ecosystems 6. Human health *Examples of key topics:* 1. State the defining characteristics that differentiate among the major taxonomic groups and organisms within these groups, and classify organisms on the basis of a variety of physical and behavioural characteristics. 2. Locate major organs in the human body, identify the components of organ systems, and compare and contrast organs and organ systems in humans and other organisms. 3. Relate the structure and function of organs and organ systems to the basic

[57] For a full account of the TIMSS 2007 science content domains, view Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., Arora, A. and Erberber, E. (2008) *TIMSS 2007 assessment frameworks*, http://timss.bc.edu/TIMSS2007/PDF/T07_AF.pdf (accessed 8 February 2010).

lungs take in oxygen, plant roots absorb water, leaves make food).	biological processes required to sustain life (sensory, digestive, skeletal and muscular, circulatory, nervous, respiratory, excretory, reproductive). 4. Explain how biological actions in response to specific external and internal changes work to maintain stable bodily conditions (eg sweating in heat, shivering in cold, increased heart rate during exercise).

Source: TIMSS 2007 report, Chapter 5 'The science curriculum', http://timss.bc.edu/TIMSS2007/PDF/T07_AF.pdf (accessed 8 February 2010).

The organisation of science topics into specific domains did not correspond to the structure of science instruction in all countries. In many countries science was taught as general science or integrated science, whereas in others science was taught as separate subjects such as biology, physics and chemistry. 8th grade students were expected to state the defining characteristics of major taxonomic groups and classify organisms according to these characteristics. They should also locate major organs and relate the structure and function of organs and organ systems to basic biological processes. TIMSS science 8th grade required students to have a foundational understanding of cells and their function, showing their ability to describe cellular make-up and to identify cell structures. Pupils had to explain how certain biological processes such as photosynthesis and respiration are necessary to sustain life. They were also to relate diversity to the survival or extinction of species in changing environments.[58]

PISA

The PISA framework focuses on what 15-year-old students can do with what they learn at school and not merely whether they have mastered curricular content. Science was the major area assessed in the 2006 study, so there was less coverage of mathematics and reading. The knowledge of science assessed in 2006 encompassed 'physical systems', 'living systems', 'Earth and space systems' and 'technology systems'. Test takers were required to engage with 'scientific enquiry' and 'scientific explanations'. The knowledge domain assessed in the mathematics section of the assessment included areas and concepts such as 'quantity', 'space and shape', 'change and relationships' and 'uncertainty'. In the reading literacy section, PISA 2006 focused on the form of reading materials: 'continuous texts',

[58] Ibid.

including different kinds of prose such as narration, exposition, argumentation, and 'non-continuous texts', including graphs, forms and lists (Table 13).

Table 13 Assessment areas, PISA 2006

PISA 2006	Science	Reading	Mathematics
Features	Possesses scientific knowledge and uses that knowledge to identify questions, acquire new knowledge, and explain scientific phenomena. Understands the features of science as a form of human knowledge and enquiry. Awareness of how science and technology shape our environment. Engages in science-related issues.	Reading literacy: the capacity to understand, use and reflect on written text to achieve a goal, to develop knowledge and as potential to participate in society. Focus on decoding and literal comprehension, interpretation and reflection underpinning the ability of reading to learn. No assessment of basic reading skills.	Mathematical literacy: the capacity to identify and understand the role that mathematics plays in the world to make well-founded judgements. The capacity to engage with mathematics to meet the needs of life as a constructive and reflective citizen. Wide functional use of mathematics; ability to recognise and formulate mathematical problems in different situations.
Knowledge domain	Physical systems Living systems Earth and space systems Technology systems Scientific enquiry Scientific explanations	The form of reading materials: ■ continuous texts (different kinds of prose: narration, exposition, argumentation) ■ non-continuous texts (graphs, forms and lists).	Relevant mathematical areas and concepts: ■ quantity ■ space and shape ■ change and relationship ■ uncertainty.

Competencies involved	Types of scientific tasks: ■ identifying scientific issues ■ explaining scientific phenomena ■ using scientific evidence.	Types of reading task: ■ retrieving information ■ interpreting texts ■ reflecting and evaluating texts.	Competency and skills needed for mathematics: ■ reproduction (simple mathematical operations) ■ connections (bringing ideas together to solve straightforward problems) ■ reflection (mathematical thinking).
Context and situation	The areas of application of science, focusing on uses in relation to: ■ health ■ natural resources ■ environment ■ hazard ■ frontiers of science ■ technology.	The use for which text is constructed: ■ private ■ public ■ occupational ■ educational contexts.	Application of mathematics. Focus on uses in relation to: ■ personal ■ educational and occupational ■ public ■ scientific settings.

In 2006 PISA included open questions requiring composition and multiple choice items. The assessment included 108 different questions at varying levels of difficulty. In many cases, students were required to construct a response in their own words. At times, students had to show their working processes.

PISA 2006 included 37 science units, comprising a total of 108 cognitive items and 31 embedded attitudinal items. The 108 science cognitive items used in the main study included 22 anchoring items from the 2003 survey. The remaining 86 items were selected from a pool of 222 newly-developed items that had been assessed in a field trial conducted in all countries in 2005, one year prior to the main study.

The distribution of the science items was as in Table 14.

Table 14 PISA 2006 science – main study items (item format by competency)

Item format	Identifying scientific issues	Explaining scientific phenomena	Using scientific evidence	Total
Multiple choice	9	22	7	38 (35%)
Complex multiple choice	10	11	8	29 (27%)
Closed constructed response	0	4	1	5 (5%)
Open constructed response	5	16	15	36 (33%)
Total	24 (22%)	53 (49%)	31 (29%)	108

Source: OECD (2009, p. 43).

According to Table 14, 38 per cent of the 108 science items in the main PISA 2006 study were multiple choice. Almost half of the items (49 per cent) asked students to explain scientific phenomena, with items on 'living systems' and on 'scientific enquiry' being the most regular in the PISA 2006 science study.

Bradshaw et al. (2007) compared the three PISA domains and methods of assessment with the national curriculum assessments and GCSEs. It seems that in the science assessments the processes of scientific enquiry and the competencies of 'identifying scientific issues', 'explaining phenomena scientifically' and 'using scientific evidence' would place students at ease because 'they are central to all science GCSE syllabuses'.[59]

[59] Ibid., p. 15.

Differences between PISA assessments and GCSE assessments would be related to the weightings given to science topics like physical science and the lack of emphasis on chemical science topics. PISA also placed a greater requirement on reading contextual information and used a greater proportion of open response items. Moreover, when comparing PISA 2003 and TIMSS 2003 mathematics content domains, Hutchison and Schagen (2006) noted that 40 per cent of PISA items involved the content 'data', as compared to only about 13 per cent items on the same content domain in TIMSS 2003. Also, looking at mathematics cognitive and competency domains, PISA 2003 dedicated only 30 per cent of its mathematics items to the reproduction of knowledge on facts and routine problems, as compared to TIMSS 2003 where the percentage of curriculum-centred items was almost 60 per cent.[60]

National curriculum assessments

The NC assessments have their content selected on the basis of its significance in the national curriculum of England. English, mathematics and science subject criteria for key stages 2 and 3 can be found in the *National curriculum assessments – regulatory framework*.[61] NC assessments are designed to test pupils' knowledge, skills and understanding as outlined in the national curriculum key stage programmes of study, which specify the different subjects to be covered. Pupils should be prepared to answer questions on any texts from the ranges specified in the national curriculum programmes of study. In England, year 5 pupils (4th grade) who are eligible to sit PIRLS should, for example, write in any of the forms covered in the key stage 2 English programme of study.[62] At key stage 3, year 9 pupils (8th grade) should answer questions on any texts specified in the key stage 3 English programme of study, which are not exemplified here. [63]For practical reasons, every

[60] See http://www.brookings.edu/gs/brown/irc2006conference/HutchisonSchagen_presentation.pdf (accessed 8 February 2010), pp. 19–21.

[61] QCA (2006) *National curriculum assessments: regulatory frame*work, QCA/06/2827.

[62] QCA (1999) *English programme of study for key stage 2 and attainment targets*, http://curriculum.qcda.gov.uk/key-stages-1-and-2/subjects/english/attainmenttargets/index.aspx?return=/key-stages-1-and-2/subjects/english/keystage2/index.aspx%3Freturn%3D/search/index.aspx%253FfldSiteSearch%253Denglish+Key+stage+2+attainment+targets+%2526btnGoSearch.x%253D39%2526btnGoSearch.y%253D14 (accessed 9 February 2010).

[63] See QCA (2007) *English programme of study for key stage 3 and attainment targets*, http://curriculum.qca.org.uk/uploads/QCA-07-3332-pEnglish3_tcm8-399.pdf?return=/key-stages-3-and-4/subjects/english/keystage3/index.aspx%3Freturn%3D/key-stages-3-and-4/subjects/index.aspx (accessed 23 February 2010).

year the NC assessments consider a small sample of learning outcomes specified in the national curriculum programmes of study. It follows that the assessments change from one year to the next, thus sampling different learning outcomes. Even when a particular learning outcome is assessed in successive years, the nature of the questions may change, which creates extra problems in terms of standard setting – a problems that does not affect the international assessments. However, norm- or cohort-referencing has been deemed a politically unacceptable alternative conceptual framework.

As Newton (2008) observed, public examinations in England were designed to measure *outcomes* representing '*end-of-course attainment* rather than examination *performance, per se*'. To clarify the point, one could say that historically, the end of key stage assessments, which cover the programmes of study prescribed by the national curriculum for that key stage, have been used to check pupil progress (alongside teacher assessment) at the end of each key stage. The level attained by a pupil is a summation of the pupil's progress. However, for 2010 the introduction of sampling for key stage 2 science will alter the uses made of the NC assessment outcomes. In line with EGA recommendations made in May 2009, sampling arrangements are being put in place to allow something to be said about national standards in science (at level 4). The science data will not be used for school or local authority accountability purposes as levels attained by pupils will not be reported. For key stage 2 English and mathematics, the arrangements remain the same as in previous years.

Despite recent changes, in contrast to PIRLS, until 2009 the national curriculum key stage 2 reading assessments were still set to follow seven assessment focuses (AFs) (Table 15). The NC assessments were designed to provide the basis for drawing inferences about reading attainment using the following level descriptors.

Table 15 Key stage 2, English subject criteria and question types, target group year 6

Reading assessment
■ AF1: Use a range of strategies, including accurate decoding of text to read for meaning. ■ AF2: Understand, describe, select or retrieve information, events or ideas from texts, and use quotation and reference to text. ■ AF3: Deduce, infer or interpret information, events or ideas from texts. ■ AF4: Identify and comment on the structure and organisation of texts, including grammatical and representational features at text level. ■ AF5: Explain and comment on writers' uses of language, including grammatical and literary features at word and sentence level. ■ AF6: Identify and comment on writers' purposes and viewpoints, and the overall effects of the text on the reader. ■ AF7: Relate texts to their social, cultural and historical contexts and literary traditions.

Source: QCA (2006, p. 12).

Along the same lines, key stage 2 mathematics questions in the NC assessments sample the full range of the AFs (Table 16) in the key stage 2 mathematics programme of study.

Table 16 Key stage 2, mathematics curriculum coverage and question types, target group year 6

■ Ma2: Number (Using and applying number; Numbers and the number system; Calculations; Solving numerical problems). ■ Ma3: Shape, space and measures (Using and applying shape, space and measures; Understanding properties of shape; Understanding properties of position and movement; Understanding measures). ■ Ma4: Handling data (Using and applying handling data; Processing, representing and interpreting data).

Source: QCA (2006, p. 17).

The mathematics curriculum leads to questions: (1) on calculation; (2) requiring the application of mathematical processes in contexts of Ma1, Ma3 and Ma4; (3) linking sections of the programme of study; (4) on mathematical reasoning; (5) that are multi-step problems; (6) drawn from both 'real-life' and mathematical contexts.[64]

In terms of science subject criteria at key stage 2, the range of questions included in the NC assessments samples the full range of AFs (Table 17) in the national curriculum key stage 2 programme of study.

Table 17 Key stage 2, science curriculum coverage and question types, target group year 6

- Sc1: Scientific enquiry – ideas and evidence; investigative skills.
- Sc2: Life processes and living things – life processes; humans and other animals; green plants; variation and classification; living things in the environment.
- SC3: Materials and their properties – grouping and classifying materials; changing materials; separating mixtures of materials.
- SC4: Physical processes – electricity; force and motion; light and sound; the Earth and beyond.

Source: QCA (2006, p. 20).

GCSEs

The regulator for England (Ofqual) and the regulators for Wales (DCELLS) and Northern Ireland (CCEA) work with teachers, awarding bodies, subject associations, higher education bodies and other interested parties to develop qualification and subject criteria for GCSEs containing specification content.[65] Exam boards publish specifications that meet the general criteria for GCSEs in the various subjects that they offer. Assessments for each qualification are carried out according to the codes of practice published by the regulatory authorities. Exam boards' GCSE specifications in each subject include:

- assessment objectives

[64] See QCA (2006, p. 18).

[65] The Ofqual GCSE criteria for English can be found at http://www.ofqual.gov.uk/files/gcse_engcriteria2002.pdf (accessed 19 February 2010).

- scheme of assessment
- specification content
- key skills
- grade descriptors.

For example, the GCSE English assessment objectives for 2007, as set by the regulator, were as shown in Table 18.

Table 18 GCSE English assessment objectives, 2007

GCSE English literature – assessment objectives	Weighting
AO1: A specification must include objectives for speaking and listening which require candidates to demonstrate their ability to: i. communicate clearly and imaginatively, structuring and sustaining their talk and adapting it to different situations, using standard English appropriately; ii. participate in discussion by both speaking and listening, judging the nature and purposes of contributions and the roles of participants; iii. adopt roles and communicate with audiences using a range of techniques.	20%
AO2: A specification must include objectives for reading which require candidates to demonstrate their ability to: i. read, with insight and engagement, making appropriate references to texts and developing and sustaining interpretations of them; ii. distinguish between fact and opinion and evaluate how information is presented; iii. follow an argument, identifying implications and recognising inconsistencies; iv. select material appropriate to their purpose, collate material from different sources, and make cross-references; v. understand and evaluate how writers use linguistic, structural and presentational devices to achieve their effects, and comment on ways language varies and changes.	20%

AO3: A specification must include objectives for writing which require candidates to demonstrate their ability to: i. communicate clearly and imaginatively, using and adapting forms for different readers and purposes; ii. organise ideas into sentences, paragraphs and whole texts using a variety of linguistic and structural features; iii. use a range of sentence structures effectively with accurate punctuation and spelling.	40%

Source: Ofqual, GCSE criteria for English,

Based on the regulator's content specifications, exam boards establish their specification content for their different qualification accreditations. It is important to note that apart from knowledge in subject matters, GCSE examinations assess key skills such as application of number, communication, information technology, improving own learning and performance, problem-solving and working with others.[66] As the regulator of external qualifications in England, Ofqual ensures the maintenance of GCE and GCSE standards for subjects over time and across awarding bodies. The Ofqual programme of standards reviews investigates examination standards to determine whether any action is needed to safeguard them.

Measurement scales

Based on regulatory documents, DCSF data and official reports published by IEA and OECD, this section explains the measurement scales that the national and international assessment frameworks use to attribute levels of achievement (IEA), levels of attainment (national curriculum assessments), proficiency levels (OECD) and grades (GCSEs) to pupils sitting these assessments and examinations. The international benchmarks, levels and grades awarded by PIRLS, TIMSS, PISA, national curriculum assessments and GCSEs can hardly be summarised in a single table. Our attempt to carry out such impossible exercise (Table 19) shows the difficulties that the media and the public might have when comparing high performance at GCSEs and at PISA. It is hard to explain what an advanced performance at TIMSS means in terms of the national curriculum assessments. One cannot say for sure that getting an F on a GCSE examination is equivalent to a level of achievement 2 in the PIRLS, TIMSS or PISA studies. However, an approximate comparison would look as follows.

[66] See Ofqual GCSE criteria for English, http://www.ofqual.gov.uk/files/gcse_engcriteria2002.pdf (accessed 8 February 2010).

Table 19 Measurement scales: PIRLS, TIMSS, PISA, NCA and GCSEs

Levels	PIRLS **Levels of achievement** International benchmarks	TIMSS **Levels of achievement** International benchmarks	PISA science **Proficiency levels** International benchmarks	NC assessments Mathematics, key stage 3 **Levels of attainment**	GCSEs **Grades**
8	–	–	–	Advanced	A*
7	–	–	–	High	A
6	–	–	Very high: 1.3% of students Lower score limit: 707.9	Intermediate	B
5	–	–	Advanced: 9.0% of students Lower score limit: 633.3	Intermediate	C
4	Advanced Lower score limit: 625	Advanced Lower score limit: 625	High: 29% of students Lower score limit: 558.7	Low	D
3	High Lower score limit: 550	High Lower score limit: 550	Intermediate: 56.7% of students Lower score limit: 484.1	–	E
2	Intermediate Lower score limit: 475	Intermediate Lower score limit: 475	Capable: 80.8% of students	–	F

			Lower score limit: 409.5		
1	Low Lower score limit: 400	Low Lower score limit: 400	Limited: 94.8% of students Lower score limit: 334.9	–	G

The GCSEs seem to award grades at higher levels than PISA. The same applies to the national curriculum assessments in comparison to PIRLS and TIMSS. Such dissimilarities could be explained in terms of greater motivation of candidates when they sit national high-stakes assessments and examinations, especially those with an impact on their career prospects. However, this section only compares the measurement scales used by PIRLS, TIMSS, PISA, national curriculum assessments and GCSES. As explained, the theoretical assumptions underpinning the measurement scales are not the same.

PIRLS

PIRLS is a norm-referenced test (NRT) in which pupils' grades depended on their position in the distribution of scores around the mean set at 500. PIRLS is a study where scores are interpreted with reference to the performance of a norm group. Pupils' performance is then compared to all other individuals taking the assessment so that one can rank pupils representing the diverse nations in the survey. IEA uses NRT design to highlight achievement differences between and among pupils to produce a dependable rank order of pupils across a continuum of achievement from high achievers to low achievers (Stiggins, 1994).

The TIMSS and PIRLS International Study Centre carried out a scale-anchoring analysis to develop descriptions of achievement at the PIRLS 2006 international benchmarks. The scale-anchoring data supplied information to describe students' performance at different points on the reading achievement scale in terms of the types of texts that they were asked to read, the types of items that they were able to answer successfully, and the quality of their answers. Such data analysis helped to identify items that discriminated between successive points on the scale (Twist et al., 2007).[67]

PIRLS measurement scales were also based on a judgemental component. A committee of reading experts examined the content of the texts and items to describe

[67] Twist et al. (2007).

pupils' comprehension skills and strategies. In contrast to the national curriculum assessments, the PIRLS reading achievement scale was designed to remain constant from assessment to assessment. However, since the countries vary in size, each country was weighted to contribute equally to the mean and standard deviation of the scales.[68]

The PIRLS international benchmarks represented the whole range of performance in all participating countries, with benchmark scores set at 'advanced' (625 points), 'high' (550 points), 'intermediate' (475 points) and 'low' (400 points) (Table 20).

Table 20 PIRLS 2006 international benchmarks

Level of achievement	Lower score limit	PIRLS 2006 international benchmarks
4	625	Advanced international benchmark – top 10% benchmark. Defined as the 90th percentile, this is the point above which the top 10% of students scored.[69] *Description* *Literary*: when reading literary texts, students can: (1) Integrate ideas across a text to provide interpretations of a character's traits, intentions and feelings, and provide full text-based support; (2) interpret figurative language; (3) begin to examine and evaluate story structure. *Informational*: when reading information texts, students can: (1) distinguish and interpret complex information from different parts of text, and provide full text-based support; (2) understand the function of organisational features; (3) integrate information across a text to sequence activities and fully justify preferences.
3	550	High international benchmark. Defined as the 75th percentile, this is the point above which the top 25 percent of students scored. *Description*

[68] The standard deviation is the statistical measure of the extent to which values are spread around the average.

[69] *PIRLS 2001 international report*, Chapter 3 'Performance at international benchmarks', http://timss.bc.edu/pirls2001i/pdf/P1_IR_Ch03.pdf (accessed 8 February 2010).

		Literary: when reading literary texts, students can: (1) locate relevant episodes and distinguish significant details embedded across the text; (2) make inferences to explain relationships between intentions, actions, events and feelings, and give text-based support; (3) recognise the use of some textual features (eg figurative language, an abstract message); (4) begin to interpret and integrate story events and character actions across the text. *Informational*: when reading information texts, students can: (1) recognise and use a variety of organisational features to locate and distinguish relevant information; (2) make inferences based on abstract or embedded information; (3) integrate information across a text to recognise main ideas and provide explanations; (4) compare and evaluate parts of a text to give a preference and a reason for it; (5) begin to understand textual elements, such as simple metaphors and author's point of view.
2	475	Intermediate international benchmark. Defined as the 50th percentile, this is the point above which the top half of the students scored. *Description* *Literary*: when reading literary texts, students can: (1) identify central events, plot sequences and relevant story details; (2) make straightforward inferences about the attributes, feelings and motivations of main characters; (3) begin to make connections across parts of the text. *Informational*: when reading information texts, students can: (1) locate and reproduce one or two pieces of information from within the text; (2) make straightforward inferences to provide information from a single part of the text; (3) use subheadings, text boxes and illustrations to locate parts of the text.
1	400	Low international benchmark. Defined as the 25th percentile, this is the point above which the top 75 percent of students scored. *Description* *Literary*: when reading literary texts, students can: (1) recognise an explicitly stated detail; (2) locate a specified part of the story and make an inference clearly suggested by the text.

		Informational: when reading information texts, students can: (1) locate and reproduce explicitly stated information that is readily accessible, eg at the beginning of the text or in a clearly defined section; (2) begin to provide a straightforward inference clearly suggested by the text.

Source: Chapter 2 'Performance at the PIRLS 2006 international benchmarks', http://timss.bc.edu/PDF/P06_IR_Ch2.pdf (accessed 8 February 2010), in Mullis et al. (2007).

TIMSS

TIMSS 2007, a norm-referenced assessment like PIRLS, also used scale-anchoring to describe student achievement at four points on the mathematics and science scales, including Advanced international benchmark (625), High international benchmark (550), Intermediate international benchmark (475) and Low international benchmark (400).

Table 21 TIMSS 2007 international benchmarks of mathematics achievement, 4th grade

Level of achievement	Lower score limit	TIMSS 2007 international benchmarks of mathematics achievement, 4th grade
4	625	Advanced international benchmark Students can apply their understanding and knowledge in a variety of relatively complex situations, and explain their reasoning. They can apply proportional reasoning in a variety of contexts. They demonstrate a developing understanding of fractions and decimals. They can select appropriate information to solve multi-step word problems. They can formulate or select a rule for a relationship. They can apply geometric knowledge of a range of two- and three-dimensional shapes in a variety of situations. Students can organise, interpret and represent data to solve problems. They can organise and draw conclusions from information, make generalisations, and solve non-routine problems. They can solve a variety of ratio, proportion and percentage problems. They can apply their knowledge of numeric and algebraic concepts and relationships. They can

		express generalisations algebraically, and model situations. They can apply their knowledge of geometry in complex problem situations. They can derive and use data from several sources to solve multi-step problems.
3	550	High international benchmark Students can apply their knowledge and understanding to solve problems. They can solve multi-step word problems involving operations with whole numbers. They can use division in a variety of problem situations. They demonstrate understanding of place value and simple fractions. They can extend patterns to find a later specified term and identify the relationship between ordered pairs. They show some basic geometric knowledge. They can interpret and use data in tables and graphs to solve problems. Students can apply their understanding and knowledge in a variety of relatively complex situations. They can relate and compute with fractions, decimals and percentages, operate with negative integers, and solve word problems involving proportions. They can work with algebraic expressions and linear equations. They use knowledge of geometric properties to solve problems, including area, volume and angles. They can interpret data in a variety of graphs, and table and solve simple problems involving probability.
2	475	Intermediate international benchmark Students can apply basic mathematical knowledge in straightforward situations. At this level they demonstrate an understanding of whole numbers. They can extend simple numeric and geometric patterns. They are familiar with a range of two-dimensional shapes. They can read and interpret different representations of the same data. Students can apply basic mathematical knowledge in straightforward situations. They can add and multiply to solve one-step word problems involving whole numbers and decimals. They can work with familiar fractions. They understand simple algebraic relationships. They demonstrate understanding of properties of triangles and basic geometric concepts. They can read and interpret graphs and tables. They recognise basic notions of likelihood.

1	400	Low international benchmark Students have some basic mathematical knowledge. They demonstrate an understanding of adding and subtracting with whole numbers. They demonstrate familiarity with triangles and informal coordinate systems. They can read information from simple bar graphs and tables. They have some knowledge of whole numbers and decimals, operations, and basic graphs.

Source: http://timss.bc.edu/timss2007/PDF/T07_M_IR_Chapter2.pdf (accessed 8 February 2010).

Countries participating in TIMSS were not always the same across the four assessments between 1995 and 2007. However, comparisons between the 2007 results and prior results were still possible because the achievement scores in each of the TIMSS surveys were placed on a scale that was not dependent on the list of participating countries in any particular year. The assessment equating and scaling details can be found in the *TIMSS 2007 technical report* (Olson et al., 2008).[70]

PISA

Like PIRLS and TIMSS, PISA is a standardised norm-referenced survey in which the norm is determined by the scores of a large group of students who sit the assessment. PISA 2006 performance scales were composed of a larger number of 'levels of proficiency' than 'levels of achievement' in PIRLS and TIMSS. Statistically derived in similar ways to PIRLS and TIMSS, the PISA scales aimed to measure a 15-year-old's ability to complete a task relating to real life. Completing tasks depended on a simple understanding of key concepts. PISA did not test subject-specific knowledge.

In 2006, about two-thirds of students sitting PISA scored between 400 and 600 points. PISA 2006 included 108 different questions at varying levels of difficulty, and each student was awarded a score based on the difficulty of questions that they could reliably tackle. Table 22 describes such levels of difficulty.

[70] Olson et al. (2008), Chapter 11 'Scaling the data from the TIMSS 2007 mathematics and science assessments', http://timss.bc.edu/TIMSS2007/PDF/T07_TR_Chapter11.pdf (accessed 8 February 2010). For further discussion on equating see Bramley, T. (2006).

Table 22 PISA 2005 student proficiency in science

Level	Lower score limit	PISA 2005 student proficiency in science
6	707.9	At level 6, students can consistently identify, explain and apply scientific knowledge and knowledge about science in a variety of complex life situations. They can link different information sources and explanations, and use evidence from those sources to justify decisions. They clearly and consistently demonstrate advanced scientific thinking and reasoning, and they demonstrate willingness to use their scientific understanding in support of solutions in unfamiliar scientific and technological situations. Students at this level can use scientific knowledge and develop arguments in support of recommendations and decisions that centre on personal, socio-economic or global situations.
5	533.3	At level 5, students can identify the scientific components of many complex life situations, apply both scientific concepts and knowledge about science to these situations, and compare, select and evaluate appropriate scientific evidence for responding to life situations. Students at this level can use well-developed inquiry abilities, link knowledge appropriately and bring critical insights to situations. They can construct explanations based on evidence and arguments based on their critical analysis.
4	558.7	At level 4, students can work effectively with situations and issues that may involve explicit phenomena requiring them to make inferences about the role of science or technology. They can select and integrate explanations from different disciplines of science or technology and link those explanations directly to aspects of life situations. Students at this level can reflect on their actions and can communicate decisions using scientific knowledge and evidence.
3	484.1	At level 3, students can identify clearly described scientific issues in a range of contexts. They can select facts and knowledge to explain phenomena and apply simple models or inquiry strategies. Students at this level can interpret and use scientific concepts from different disciplines and can apply them directly. They can develop short statements using facts

		and make decisions based on scientific knowledge.
2	409.5	At level 2, students have adequate scientific knowledge to provide possible explanations in familiar contexts or draw conclusions based on simple investigations. They are capable of direct reasoning and making literal interpretations of the results of scientific inquiry or technological problem-solving.
1	334.9	At level 1, students have such a limited scientific knowledge that it can be applied to only a few, familiar situations. They can present scientific explanations that are obvious and follow explicitly from given evidence.

Source: http://www.oecd.org/dataoecd/30/17/39703267.pdf (accessed 8 February 2010).

PISA 2006 proficiency levels in mathematics (Table 23) were the same as those established for the same subject when it was the major area of assessment in PISA 2003. In mathematics there were six levels of proficiency, with cut-scores being slightly different from those in the science scale.

Table 23 PISA 2005 student proficiency in mathematics

Level	Lower score limit	PISA 2005 student proficiency in mathematics
6	669.3	At level 6, students can conceptualise, generalise and utilise information based on their investigations and modelling of complex problem situations. They can link different information sources and representations, and flexibly translate among them. Students at this level are capable of advanced mathematical thinking and reasoning. They can apply this insight and understanding along with a mastery of symbolic and formal mathematical operations and relationships to develop new approaches and strategies for attacking novel situations.
5	607.0	At level 5, students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well-developed thinking and reasoning skills, appropriate linked representations, symbolic

		and formal characterisations, and insight pertaining to these situations.
4	544.7	At level 4, students can work effectively with explicit models for complex concrete situations that may involve constraints or call for making assumptions. They can select and integrate different representations, including symbolic ones, linking them directly to aspects of real-world situations. Students at this level can utilise well-developed skills and reason flexibly, with some insight, in these contexts. They can construct and communicate explanations and arguments based on their interpretations, arguments and actions.
3	482.4	At level 3, students can execute clearly described procedures, including those that require sequential decisions. They can select and apply simple problem-solving strategies. Students at this level can interpret and use representations based on different information sources, and reason directly from them. They can develop short communications reporting their interpretations, results and reasoning.
2	420.1	At level 2, students can interpret and recognise situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures or conventions. They are capable of direct reasoning and making literal interpretations of the results.
1	375.8	At level 1, students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are obvious and follow immediately from the given stimuli.

Source: http://www.oecd.org/dataoecd/30/17/39703267.pdf (accessed 8 February 2010), p. 312.

PISA 2006 proficiency levels in reading (Table 24) had one level less than the science and mathematics scales.

Table 24 PISA 2005 student proficiency in reading

Level	Lower score limit	PISA 2005 student proficiency in reading
5	625.6	Students can locate and possibly sequence or combine multiple pieces of deeply embedded information, some of which may be outside the main body of the text. They can infer which information in the text is relevant to the task, and deal with highly plausible and/or extensive competing information. They can either construe the meaning of nuanced language or demonstrate a full and detailed understanding of a text. They can critically evaluate or hypothesise, drawing on specialised knowledge. They can deal with concepts that are contrary to expectations and draw on a deep understanding of long or complex texts. In continuous texts, students can analyse texts whose discourse structure is not obvious or clearly marked, in order to discern the relationship of specific parts of the text to its implicit theme or intention. In non-continuous texts, students can identify patterns among many pieces of information presented in a display that may be long and detailed, sometimes by referring to information external to the display.
4	552.9	Students can locate and possibly sequence or combine multiple pieces of embedded information, each of which may need to meet multiple criteria, in a text with familiar context or form. They can infer which information in the text is relevant to the task. They can use a high level of text-based inference to understand and apply categories in an unfamiliar context, and to construe the meaning of a section of text by taking into account the text as a whole. They can deal with ambiguities, ideas that are contrary to expectation and ideas that are negatively worded. They can use formal or public knowledge to hypothesise about or critically evaluate a text. They can show accurate understanding of long or complex texts. In continuous texts, students can follow linguistic or thematic links over several paragraphs, often in the absence of clear discourse markers, in order to locate, interpret or evaluate embedded information or to infer psychological or metaphysical meaning. In non-continuous texts, students can scan a long,

		detailed text in order to find relevant information.
3	480.2	Students can locate, and in some cases recognise, the relationship between pieces of information, each of which may need to meet multiple criteria. They can deal with prominent competing information. They can integrate several parts of a text in order to identify a main idea, understand a relationship or construe the meaning of a word or phrase. They can compare, contrast or categorise, taking many criteria into account, and deal with competing information. They can make connections or comparisons, give explanations, or evaluate a feature of text. They can demonstrate a detailed understanding of the text in relation to familiar, everyday knowledge, or draw on less common knowledge. In continuous texts, students can use conventions of text organisation, where present, and follow implicit or explicit logical links – such as cause and effect relationships across sentences or paragraphs – in order to locate, interpret or evaluate information. In non-continuous texts, students can consider one display in the light of a second, separate document or display.
2	407.5	Students can locate one or more pieces of information, each of which may be required to meet multiple criteria. They can deal with competing information. They can identify the main idea in a text, understand relationships, form or apply simple categories, or construe meaning within a limited part of the text when the information is not prominent and low-level inferences are required. They can make a comparison or connections between the text and outside knowledge, or explain a feature of the text by drawing on personal experience and attitudes. In continuous texts, students can follow logical and linguistic connections within a paragraph in order to locate or interpret information, or synthesise information across texts or parts of a text in order to infer the author's purpose. In non-continuous texts, students demonstrate a grasp of the underlying structure of a visual display such as a simple tree diagram or table, or combine two pieces of information from a graph or table.
1	334.8	Students can locate one or more independent pieces of explicitly stated information, typically meeting a single criterion, with little or no competing information in the text. They can

		recognise the main theme or author's purpose in a text about a familiar topic, when the required information in the text is prominent. They can make a simple connection between information in the text and common, everyday knowledge. In continuous texts, students can use redundancy, paragraph headings or common print conventions to form an impression of the main idea of the text, or to locate information stated explicitly within a short section of text. In non-continuous texts, students can focus on discrete pieces of information, usually within a single display such as a simple map, a line graph or a bar graph that presents only a small amount of information in a straightforward way and in which most of the verbal text is limited to a small number of words or phrases.

Source: http://www.oecd.org/dataoecd/30/17/39703267.pdf (accessed 8 February 2010), pp. 292–3.

National curriculum assessments

In contrast to the international surveys, the NC assessments are criterion-referenced assessments that measure pupils' knowledge and understanding against specific standards over time. A pupil's performance in the assessment is not compared to the performance of other pupils, as it is the case with norm-referenced assessments like PIRLS, TIMSS and PISA. In the NC assessments it is possible for all pupils to earn the highest grade, if all meet the established criteria. The England regulator of qualifications and exams publishes *level descriptors* or *attainments targets* that are derived from specialist judgement and used to allocate pupils to different levels of attainment in the national curriculum.

When the regulator describes *subject criteria* for key stages 2 and 3 NC assessments, it also establishes the levels of achievement to be assessed. According to the QCA regulatory framework for key stages 1 to 3 (key stage 3 revised in 2006), the key stage 2 English assessments (established in 1999) must target the year 6 group, being designed to assess 10- to 11-year-old pupils' performance at levels 2–5. The year 6 group is expected to attain level 4 at the end of key stage 2.

In England pupils are judged before they sit the assessment, which will identify their progress in the scale of attainment. This is not the case with the international assessments. As shown in Table 25, the national curriculum attainment targets describe eight levels of progressive difficulty, and an explanation of performance above level 8. Each level descriptor spells out the type and range of performance that pupils working at that level should demonstrate. NC assessments *attainment targets* for key stage 2 English refer to (a) 'Speaking and listening', (b) 'Reading' and

(c) 'Writing', but for the sake of space Table 25 gives an example of just 'Reading' attainment targets.

Table 25 Key stage 2 English (En2) 'Reading' attainment targets (year 6)

Level	Attainment targets
Exceptional performance	Pupils confidently sustain their responses to a demanding range of texts, developing their ideas and referring in detail to aspects of language, structure and presentation. They make apt and careful comparison between texts, including consideration of audience, purpose and form. They identify and analyse argument, opinion and alternative interpretations, making cross-references where appropriate.
8	Pupils' responses are shown in their appreciation of, and comment on, a range of texts, and they evaluate how authors achieve their effects through the use of linguistic, structural and presentational devices. They select and analyse information and ideas, and comment on how these are conveyed in different texts.
7	Pupils show understanding of the ways in which meaning and information are conveyed in a range of texts. They articulate personal and critical responses to poems, plays and novels, showing awareness of their thematic, structural and linguistic features. They select and synthesise a range of information from a variety of sources.
6	In reading and discussing a range of texts, pupils identify different layers of meaning and comment on their significance and effect. They give personal responses to literary texts, referring to aspects of language, structure and themes in justifying their views. They summarise a range of information from different sources.
5	Pupils show understanding of a range of texts, selecting essential points and using inference and deduction where appropriate. In their responses, they identify key features, themes and characters, and select sentences, phrases and relevant information to support their views. They retrieve and collate information from a range of sources.
4	In responding to a range of texts, pupils show understanding of significant ideas, themes, events and characters, beginning to use inference and deduction. They refer to the text when explaining their

	views. They locate and use ideas and information.
3	Pupils read a range of texts fluently and accurately. They read independently, using strategies appropriately to establish meaning. In responding to fiction and non-fiction they show understanding of the main points and express preferences. They use their knowledge of the alphabet to locate books and find information.
2	Pupils' reading of simple texts shows understanding and is generally accurate. They express opinions about major events or ideas in stories, poems and nonfiction. They use more than one strategy, such as phonic, graphic, syntactic and contextual, in reading unfamiliar words and establishing meaning.
1	Pupils recognise familiar words in simple texts. They use their knowledge of letters and sound-symbol relationships in order to read words and establish meaning when reading aloud. In these activities they sometimes require support. They express their response to poems, stories and non-fiction by identifying aspects that they like.

Source: QCA (1999).

The above descriptors are helpful in explaining the PIRLS 2006 international benchmarks from a national curriculum perspective. Whereas 10- to 11-year-old children in England attaining English level 8 can show their appreciation of, and comment on, a range of texts, being able to evaluate how authors achieve their effects through the use of linguistic, structural and presentational devices, the top 10 per cent of children scoring 625 points or more in PIRLS assessments can: (1) integrate ideas across a text to provide interpretations of a character's traits, intentions and feelings, and provide full text-based support; (2) interpret figurative language; (3) begin to examine and evaluate story structure. That is, the key stage 2 national curriculum attainment targets for reading are far more ambitious than PIRLS international benchmarks for children in the same age group. Yet low international assessments demand may be a problem affecting the international assessing of English language alone.

In 2007, the *attainment targets* for key stage 3 mathematics were: (a) mathematical processes and applications; (b) number and algebra; (c) geometry and measures; (d) handling data. These targets were assessed on levels 3 to 8, with written papers for tiers 3 to 5, 4 to 6, 5 to7 and 6 to8. Table 26 displays examples of some levels of attainment on the subject matter, excluding level 3 and exceptional performance.

Table 26 Key stage 3 'Number and algebra' mathematics attainment targets (year 9)

Level	Attainment targets
8	Pupils solve problems that involve calculating with powers, roots and numbers expressed in standard form. They choose to use fractions or percentages to solve problems involving repeated proportional changes or the calculation of the original quantity given the result of a proportional change. They evaluate algebraic formulae or calculate one variable given the others, substituting fractions, decimals and negative numbers. They manipulate algebraic formulae, equations and expressions, finding common factors and multiplying two linear expressions. They solve inequalities in two variables. They sketch and interpret graphs of linear, quadratic, cubic and reciprocal functions, and graphs that model real situations.
7	When making estimates, pupils round to one significant figure and multiply and divide mentally. They understand the effects of multiplying and dividing by numbers between 0 and 1. They solve numerical problems involving multiplication and division with numbers of any size, using a calculator efficiently and appropriately. They understand and use proportional changes, calculating the result of any proportional change using only multiplicative methods. They find and describe in symbols the next term or *n*th term of a sequence where the rule is quadratic. They use algebraic and graphical methods to solve simultaneous linear equations in two variables.
6	Pupils order and approximate decimals when solving numerical problems and equations, using trial and improvement methods. They evaluate one number as a fraction or percentage of another. They understand and use the equivalences between fractions, decimals and percentages, and calculate using ratios in appropriate situations. They add and subtract fractions by writing them with a common denominator. They find and describe in words the rule for the next term or *n*th term of a sequence where the rule is linear. They formulate and solve linear equations with whole-number coefficients. They represent mappings expressed algebraically, and use Cartesian coordinates for graphical representation interpreting general features.
5	Pupils use their understanding of place value to multiply and divide whole numbers and decimals. They order, add and subtract negative numbers in context. They use all four operations with decimals to two places. They solve simple problems involving ratio and direct

	proportion. They calculate fractional or percentage parts of quantities and measurements, using a calculator where appropriate. They construct, express in symbolic form and use simple formulae involving one or two operations. They use brackets appropriately. They use and interpret coordinates in all four quadrants.
4	Pupils use their understanding of place value to multiply and divide whole numbers by 10 or 100. When solving number problems, they use a range of mental methods of computation with the four operations, including mental recall of multiplication facts up to 10 times t 10 and quick derivation of corresponding division facts. They use efficient written methods of addition and subtraction, and of short multiplication and division. They recognise approximate proportions of a whole, and use simple fractions and percentages to describe these. They begin to use simple formulae expressed in words.

Source: QCA (2007) Mathematics programme of study for key stage 3 and attainment targets, pp. 148–53.

The science programme of study also described levels of attainment 4 to 8 for the following attainment targets: attainment target 1 'How science works'; attainment target 2 'Organisms, their behaviours and the environment'; attainment target 3 'Materials and their properties and the Earth'; attainment target 4 'Energy, forces and space'.[71]

In summary, the NC assessments interpret candidates' performance in relation to predetermined criteria. Emphasis is on attainment of objectives rather than on candidates' scores as a reflection of their ranking within the group. Criterion-referenced tests (CRTs) like the NC assessments establish what pupils can do and what they know, not how they compare to others. CRTs report how well pupils are doing relative to a predetermined performance level on a specified set of educational goals or outcomes included in the national curriculum. Policy makers opted to use a CRT in England for several reasons, including their need to know how schools are teaching the national curriculum (Bond, L.A.1996).

The English regulator establishes the standards of assessments and suggests models for maintaining the standards of the NC assessments. One of these models, the *level-setting* meeting, must consider the thresholds between levels of attainment

[71] See QCA (2007) *Science: programme of study for key stage 3 and attainment targets*, http://curriculum.qca.org.uk/uploads/QCA-07-3344-p_Science_KS3_tcm8-413.pdf?return=/key-stages-3-and-4/subjects/science/keystage3/index.aspx%3Freturn%3D/key-stages-3-and-4/subjects/science/index.aspx (accessed 8 February 2010).

and associated ranges. Level-setting meetings aim to reach consensus on a draft level threshold mark or zone and script scrutiny range.[72] This explains why the national curriculum attainment targets do not include lower score limits or cut-scores like PIRLS, TIMSS and PISA do. Professional judgement is more relevant in the NC assessments than in the international ones.

GCSEs

Like the national curriculum assessments, grade descriptions for GCSEs are provided by the regulatory authority. Such descriptors give a general indication of the standards of achievement likely to have been shown by candidates awarded particular grades. Grade descriptions must be interpreted in relation to the content in the specification or defined in the subject criteria (Table 27). Descriptors are not designed to define examinations content that is set by exam boards. Both subject criteria and grade descriptors are intended to help to ensure consistent and comparable standards in the same subject area across awarding bodies.

QCDA publishes subject criteria for all GCSE subjects[73] describing the knowledge, understanding, skills and schemes of assessment. The subject criteria provide the framework within which an exam board can create the detail of the specification (Table 27).[74] The grade awarded to a student will depend on the extent to which the candidate has met the assessment objectives overall. Deficiency in some aspects of a candidate's performance in the assessment may be balanced by better performances in other areas of the exam.

[72] The recommended format for a draft level-setting format can be found in QCA (2007) *The national curriculum assessments Code of Practice, key stages 1–3*.

[73] View for example subject criteria for GCSE in engineering at http://www.qca.org.uk/qca_5547.aspx (accessed 22 February 2010).

[74] View GCSE qualification and subject criteria at http://www.ofqual.gov.uk/743.aspx (accessed 8 February 2010).

Table 27 GCSE English grade descriptors

GCSE English[75] Grade	Grade descriptors
A	Candidates select suitable styles and registers of spoken English for a range of situations and contexts, showing assured use of standard English where appropriate. They confidently vary sentence structures and choose from a broad repertoire of vocabulary to express information, ideas and feelings in an engaging manner. They initiate conversations and demonstrate sensitive listening through contributions that sustain and develop discussion. They recognise and fulfil the demands of different roles, whether in formal settings or creative activities. Candidates respond personally and persuasively to a variety of texts, developing interpretations and evaluating how details of language, grammar, structure and presentation engage and affect the reader. They identify and discuss writers' perspectives in narrative, argument, explanation or analysis. They choose apt quotations and make telling comparisons and cross-references that illuminate the purpose and meanings of texts, explaining the impact of their social, cultural and historical contexts where appropriate. Candidates' writing shows confident, assured control of a range of forms and styles appropriate to task and purpose. Texts engage and hold the reader's interest through logical argument, persuasive force or creative delight. Linguistic and structural features are used skilfully to sequence texts and achieve coherence. A wide range of accurate sentence structures ensures clarity; choices of vocabulary, punctuation and spelling are ambitious, imaginative and correct.
C	Candidates adapt their talk to the demands of different situations and contexts. They recognise when standard English is required and use it confidently. They use different sentence structures and select vocabulary so that information, ideas and feelings are communicated clearly and the listener's interest is engaged. They explain and evaluate how they and others use and adapt spoken language for specific purposes. Through careful listening and by developing their

[75] See Ofqual *GCSE criteria for English*, http://www.ofqual.gov.uk/files/gcse_engcriteria2002.pdf (accessed 8 February 2010).

	own and others' ideas, they make significant contributions to discussion and participate effectively in creative activities. Candidates understand and demonstrate how meaning and information are conveyed in a range of texts. They make personal and critical responses, referring to specific aspects of language, grammar, structure and presentational devices to justify their views. They successfully compare and cross-reference aspects of texts and explain convincingly how they may vary in purpose and how they achieve different effects. They comment on how social, cultural and historical contexts affect readers' responses to texts. Candidates' writing shows successful adaptation of form and style to different tasks and for various purposes. They use a range of sentence structures and varied vocabulary to create different effects and engage the reader's interest. Paragraphing is used effectively to make the sequence of events or development of ideas coherent and clear to the reader. Sentence structures are varied; punctuation and spelling are accurate and sometimes bold.
F	Candidates talk confidently in familiar situations, showing some awareness of purpose and of listeners' needs. They convey information, develop ideas and describe feelings clearly, using the main features of standard English as appropriate. They listen with concentration and make relevant responses to others' ideas and opinions. They show some awareness of how they and others use and adapt spoken language for specific purposes. In formal and creative activities, they attempt to meet the demands of different roles. Candidates describe the main ideas, themes or arguments in a range of texts, and refer to specific aspects or details when justifying their views. They make simple comparisons and cross-references that show some awareness of how texts achieve their effects through writers' use of linguistic, grammatical, structural and presentational devices. They are aware that some features of texts relate to their specific social, cultural and historical contexts. Candidates' writing shows some adaptation of form and style for different tasks and purposes. It communicates simply and clearly with the reader. Sentences sequence events or ideas logically; vocabulary is sometimes chosen for variety and interest. Paragraphing is straightforward but effective; the structure of sentences, including some that are complex, is usually correct. Spelling and basic punctuation are

	mostly accurate.

Source: Ofqual (2009c).

Curriculum match

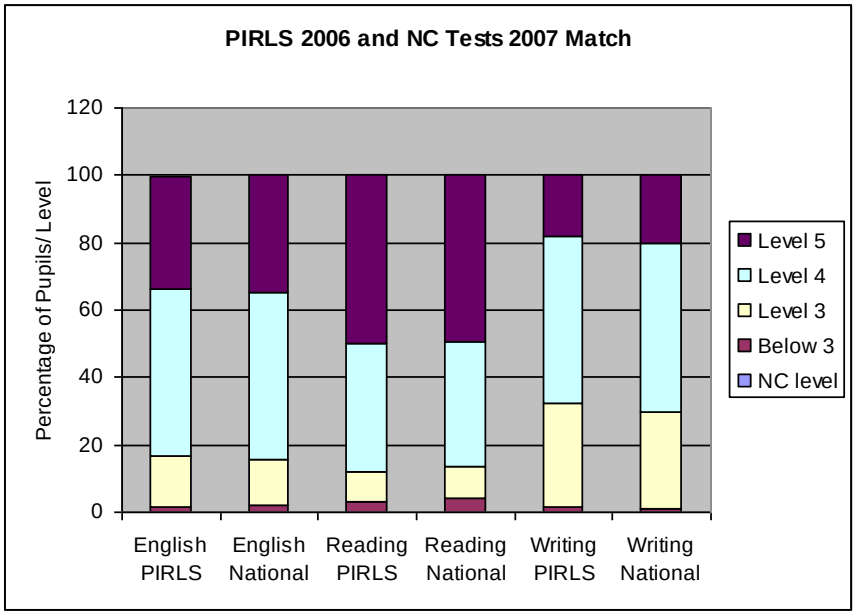
This section explores the extent to which certain national and international assessments aiming at the same age groups and focusing on the same topics have overlapping areas. There certainly is a degree of match between the underlying focuses of PIRLS, TIMSS and PISA assessments and the theories, processes and cognitive demands (complexity, resources, abstractness, task strategy and response strategy) specified in the national curriculum assessments and GCSEs. The comparability[76] of assessment results depends on whether the assessments measure the same 'construct' or assumed ability, which cannot be directly observed or measured. There are certainly different constructs underlying reading ability, scientific ability and mathematics ability in general. In this sense, diverse mathematics assessments assessing different constructs would not say the same things about the same pupils taking the assessment. However, in many respects PIRLS, TIMSS and PISA attempt to measure the similar constructs that underlie the mathematical, scientific and reading ability measured by the national curriculum assessments and GCSEs.

PIRLS

According to the NFER PIRLS 2006 national report, there was a good match of attainment and performance between the national population sitting key stage 2 assessments and the PIRLS sample. That is, the English sample was representative. There was only a slight difference between the distribution of levels achieved in reading and in writing of the PIRLS pupils when compared with the national distribution (Chart 1). As Twist et al. (2007, p. 134) observed, the distribution of national curriculum assessment levels attained in 'reading' was slightly slanted towards the upper end of achievement, and in 'writing' towards the lower end when compared with PIRLS results.

[76] For a study on the comparability of assessments see Massey, A., Green, S., Dexter, T. and Hamnett, L. (2003) *Comparability of national tests over time: KS1, KS2 and KS3 standards between 1996 and 2001.*

Chart 1 National curriculum levels achieved by PIRLS sample in key stage 2 English assessment in 2007 compared to national distribution



Source: Twist et al. (2007, p. 134).

Sainsbury (2003, p. 32)[77] previously noted that children in England following the national curriculum had been well prepared for the demands of the PIRLS literacy study because there were large overlapping areas. Yet the correlation between pupils' scores on PIRLS 2001 reading and on the national curriculum key stage 2 reading assessment one year later (0.77)[78] was not particularly high but at the same level of the usual correlation for two assessments taken by similar cohorts.

In 2001, the national curriculum required both literature and factual reading. Children were taught to use inference, to formulate opinions and to analyse what they had read. When the PIRLS children went on to take their national curriculum key stage 2 assessments in 2002, some of the questions in the national curriculum assessments were similar to those in PIRLS and a few were more demanding than PIRLS items. PIRLS 2001 items on reading process and some of the key stage 2 assessments assessment focuses clearly matched (Table 28). Yet PIRLS covered just a fraction of the 2002 English curriculum, which demanded pupils to comment on the structure and organisation of texts, explaining writers' use of language, purposes and viewpoints, and the effect of the text on the reader. Pupils sitting the national

[77] Sainsbury (2003, p.32) 'Reading Literacy Tests' in Twist et. al (2003) *Reading All over the World, NFER/ Department for Education and Skills.*

[78] Sainsbury (2003, p.4) Background to PIRLS 2001 in Twist et. al (2003) *Reading All over the World, NFER/ Department for Education and Skills.*

curriculum assessments were expected to do more as they had to relate texts to their social, cultural and historical contexts and literary traditions.[79]

Table 28 Processes of comprehension, PIRLS 2001 and key stage 2 national curriculum assessments 2002

PIRLS – processes of comprehension	National curriculum assessments – English assessment focuses (AFs)
Focus on and retrieve explicitly stated information and ideas	AF2: understand, describe, select or retrieve information, events or ideas from texts, and use quotation and reference to text.
Make straightforward inferences	AF3: deduce, infer or interpret information, events or ideas from texts.
Interpret and integrate ideas	AF3: deduce, infer or interpret information, events or ideas from texts.
Examine and evaluate content	AF4: identify and comment on language and textual elements of the structure and organisation of texts. AF5: explain and comment on writers' use of language, including grammatical and literary features at word and sentence level. AF6: identify and comment on writers' purposes and viewpoints, and the effect of the text on the reader. AF7: relate texts to their social, cultural and historical contexts and literary traditions.

Source: Sainsbury (2003, p. 44, Figure 4.13, PIRLS processes and national curriculum assessment focuses).

[79] For a comparison of the core primary curriculum in England with those of other high performing countries see also Ruddock, G. and Sainsbury, M. with Clausen-May, T., Vappula, H., Mason, K., Patterson, E.W., Pyle, K., Kispal, A., Siddiqui, R., McNaughton, S. and Rees, F. (2008)

TIMSS

IEA have published a series of test-curriculum matching analysis (TCMA) reports to show the extent to which students from different countries are familiar with TIMSS items. Additionally, TCMA shows how student performance for individual countries varies when based only on the assessment questions that are judged to be relevant to their own curricula. An item was to be deemed appropriate if it was in the intended curriculum for more than 50 per cent of the students. Table 29 illustrates the match between TIMSS items and typical items in the national curriculum assessments from 1995 to 2007. The match ranged from 70 per cent (TIMSS science 2007 – 4th grade, year 5) to 97 per cent (TIMSS mathematics 2003 – 8th grade, key stage 3, year 9).

It is noteworthy that when TIMSS matched the national curriculum assessments well in 2003 (Mullis et al, 2004, pp.400-402)[80] (Table 29), year 9 students from England performed relatively less well in mathematics.[81] Therefore in practice curriculum matching does not necessarily guarantee fairness and better performance in the international assessments because there are other background factors influencing performance, such as the training of teachers in mathematics, the influence of textbooks, the course content that was actually taught, and the effectiveness of different instructional practices.

[80] Mullis, I.V.S., Martin, M.O., Gonzales, E.J. and Chrostowski, S.J. (2004) *TIMSS 2003 international mathematics report*, The Test Curriculum Matching Analysis, Appendix C, http://timss.bc.edu/PDF/t03_download/T03INTLMATRPT.pdf (accessed 9 February 2010), pp. 400–2.

[81] Mullis, I.V.S., Martin, M.O., Gonzales, E.J. and Chrostowski, S.J. *TIMSS 2003 international mathematics report*, http://timss.bc.edu/PDF/t03_download/T03_M_Chap1.pdf (accessed 9 February 2010), p. 32.

Table 29 Assessment–curriculum matching analysis: TIMSS and England's national curriculum, 1995–2007

Year	1995 Maths Grade 8 (Year 9)	1995 Science Grade 8	2003 Maths Grade 4	2003 Science Grade 4	**2003 Maths Grade 8**	2003 Science Grade 8	2007 Maths Grade 4	2007 Science Grade 4 (Year 5)	2007 Maths Grade 8	2007 Science Grade 8
Items NC from England	130	124	161 (96%)	134 (81%)	208 (97%)	199 (96%)	184 (98%)	131 (70%)	226 (96%)	199 (86%)
Total items or score points at TIMSS	?	?	166	165	213	206	188	189	236	231
Correct England	53%	61%	61%	63%	46%	55%	61%	59%	52%	54%
Correct international average	55%	55%	53%	55%	41%	43%	495	50%	40%	41%
Standard error England	0.7	0.6	0.8	0.7	1.2	0.9	0.7	0.6	1.2	1.0
Standard error international average	0.9	0.7	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Average score in the assessments		533	531	540	498	544	541	542	513	542

Source: IEA, *TIMSS international mathematics reports:* The Test Curriculum Matching Analysis 1995,[82] 2003[83] and 2007.[84]

[82] IEA (1996) Beaton, A.; Mullis, I.V.S.; Martin, M.O.; Gonzalez, E.J.; Kelly, D.L.; Smith, T.A. *Mathematics Achievement in the Middle School Years*, IEA, Appendix B. http://timss.bc.edu/timss1995i/MathB.html

[83] IEA (2004) Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Chrostowski, S.J. (2004), Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. TIMSS 2003 International Mathematics *report*, Appendix C http://timss.bc.edu/PDF/t03_download/T03_M_AppC.pdf

[84] IEA (2008) Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston

Over the years, TCMA challenged ingrained beliefs that TIMSS participants could benefit from including their national curriculum questions in the international assessments. More importantly, one assumed that by omitting undesirable items for one country, TIMSS would enhance the results for that country. Now it is clear that omitting undesirable items for one country will also improve the results for all other countries, so that the overall pattern of results is largely unaffected because items that are difficult for England's students tend to be difficult for students in other countries as well.

Mathematics

An analysis of average percentage correct answers in TIMSS 2007 mathematics 4th grade showed how pupils form England scored 61 per cent correct answers on assessment items included by England but also scored very similar percentages on questions included by other countries. Considering TIMSS 2007 4th grade, the selection of international items also did not have a major effect on the average of correct answers given by participants from England and from other countries.[85]

The TIMSS 2007 mathematics report compared countries according to yearly hours of mathematics instruction, revealing that at 4th grade (year 5), pupils in England received 183 hours of mathematics instruction at school in 2007, whereas the international average of instructional time spent on mathematics was 144 hours. However, at 8th grade (year 9), pupils in England were at disadvantage because they received only 113 hours of mathematics instruction at school, compared to 120 hours of international average mathematics instructional time which placed pupils from other countries at some advantage.[86] Table 30 shows that 85 per cent of all year 5 pupils in England were taught all 35 TIMSS mathematics topics tested in the 4th grade assessment, whereas the international average was 66 per cent.

College. *TIMSS 2007International Mathematics report*, Appendix C
http://timss.bc.edu/TIMSS2007/PDF/T07_M_IR_AppendixC.pdf

[85] IEA (2008) Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. *TIMSS 2007International Mathematics report*, Appendix C
http://timss.bc.edu/TIMSS2007/PDF/T07_M_IR_AppendixC.pdf

[86] Source: TIMSS 2007 Chapter 5 'The mathematics curriculum',
http://timss.bc.edu/TIMSS2007/PDF/T07_M_IR_Chapter5.pdf (accessed 9 February 2010), pp. 194–5.

Table 30 Percentage of students taught the TIMSS mathematics topics (4th grade)

Country	All mathematics (35 topics)	Number (19 topics)	Geometric shapes and measures (11 topics)	Data display (5 topics)
England KS2 (year 5), 4th grade	85	85	88	83
International average, 4th grade	66	70	64	64

England's percentage of students taught TIMSS mathematics topics was clearly higher than the international average at 4th grade.[87] The same applies throughout the 8th grade (Table 31) but the percentage of students being taught TIMSS mathematics topics was generally lower than benchmarking jurisdictions such as Massachusetts (USA) and Minnesota (USA), with the exception of geometry.[88]

Table 31 **Percentage of students taught the TIMSS mathematics topics (8th grade)**

Country	All mathematics (39 topics)	Number (10 topics)	Algebra (8 topics)	Geometry (14 topics)	Data and chance (7 topics)
England KS3 (year 9), 8th grade	86	97	84	83	81
International average, 8th grade	72	95	73	71	47

Source: http://timss.bc.edu/TIMSS2007/PDF/T07_M_IR_Chapter5.pdf, p.202-3

[87] Ibid., p. 204.

[88] Ibid., p. 205.

Science

The TIMSS 2007 science report revealed whether the TIMSS science topics were included in the intended curriculum taught at schools in the 57 TIMSS countries. The percentages of students taught TIMSS science topics in England are displayed in Tables 32 and 33.

Table 32 Percentage of students taught the TIMSS science topics (4th grade)

Country	Life science	Physical science	Earth science	Other
England KS3 (year 9), 4th grade	37	36	24	3
International average, 4th grade	40	25	24	10

Source: Martin, M.O., Mullis, I.V.S. and Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C. and Galia, J.) (2008) http://timss.bc.edu/TIMSS2007/PDF/T07_S_IR_Chapter5.pdf (accessed 9 February 2010), p. 210.

Table 33 Percentage of students taught the TIMSS science topics (8th grade)

Country	Biology	Chemistry	Physics	Earth science	Other
England KS3, 8th grade	29	29	28	10	4
International average, 8th grade	28	24	27	16	6

Source: Martin, M.O., Mullis, I.V.S. and Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C. and Galia, J.) (2008) http://timss.bc.edu/TIMSS2007/PDF/T07_S_IR_Chapter5.pdf (accessed 9 February 2010), p. 211.

In England, 27 out of 35 TIMSS 2007 science topics were included in the national curriculum key stage 2. However, 8 TIMSS topics were not in the national curriculum.[89] At 8th grade, the majority of TIMSS science assessment topics (40 out

[89] Martin, M.O., Mullis, I.V.S., and Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. TIMSS 2007

of 46) were included in the key stage 3 curriculum, and only 6 TIMSS topics were not in the national curriculum of England.[90]

There are enough data to prove that the national curriculum framework and TIMSS 2007 framework have tended to be similar.[91] But there were some differences, like the presence of tiering arrangements and mental mathematics in the national curriculum assessments. Moreover, in 2007 the mathematics national curriculum assessment took 1 hour and 50 minutes in total, being far more complex than TIMSS assessments, which last 45 minutes with 215 items.

PISA

Studies on the overlapping areas between the PISA surveys and the national curriculum of England gained insight from Smithers (2004). The author noted that the results for mathematics and science TIMSS 1999 and PISA 2000 from the same cohort of pupils assessed one year apart were compared and found to be very different.[92] The discrepancies were attributable mainly to differences in the aims of the assessments, the types of questions, target populations, and response rates. Smithers (2004) suggested that the PISA programme did not demonstrate that its literacy surveys were indeed measures of 'knowledge and skills for life'. PISA's mathematics and science surveys were deemed to be assessments of reading centred on elementary mathematical and scientific concepts, and were too different from the construct assessed in GCSE examinations.

Adopting a different research strategy, Mullis et al. (2006, p. 101) compared PIRLS and PISA frameworks for the assessment of reading literacy, demonstrating how two different international consensus-building processes could result in somewhat similar approaches to assessment in terms of the construct being measured. The similarities and differences between PIRLS and PISA seemed 'developmentally appropriate'.[93] Both viewed reading as an interactive, constructive process.

International Science Report. Chapter 5. The Science Curriculum. http://timss.bc.edu/TIMSS2007/sciencereport.html, p.214.

[90]Martin, M.O. et al (2008), p. 216.

[91] Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. Appendix C, http://timss.bc.edu/TIMSS2007/PDF/T07_M_IR_AppendixC.pdf*

[92] Smithers, A. (2004) England's education: what can be learned by comparing countries?', University of Liverpool, http://www.suttontrust.com/reports/pisa_publication.doc (accessed 9 February 2010).

[93] Mullis et al. (2006).

Taking a different approach, Bradshaw et al. (2007) warned that PISA 2006 reading literacy and English GCSEs did not measure exactly the same construct since there was a limited overlap between the notion of 'reading literacy' in PISA 2006 – seeking to measure a student's ability to understand, use and reflect on a variety of written texts and different contexts – and the attainment targets for English in the national curriculum. English GCSE papers also aimed to assess the skills of 'speaking', 'listening' and 'writing' using a wider range of literary texts, which include drama, poetry and prose fiction. In science the content areas of 'Earth and space', 'living systems', 'physical systems' and 'technology systems' would also be quite familiar to GCSE students. In terms of the processes of scientific enquiry and competencies such as 'identifying scientific issues', 'explaining phenomena scientifically' and 'using scientific evidence', these were crucial to all GCSE syllabuses.[94] A QCA study entitled *PISA 2006: a first summary of the results for England*[95] had agreed that overall there was a good match between the content areas and processes of science assessed in PISA 2006 and those specified in the national curriculum for science.

In 2008 the DCSF conducted an internal study with NFER matching PISA 2003 and PISA 2006 pupils' records, and matching these with the National Pupil Database (NPD).[96] The study identified matching areas between PISA and the national curriculum, and a possible convergence between the assessments could be leading the government to view the OECD survey as the favourite option to monitor the extent to which England is getting closer to world-class education systems.

PISA surveys are known to have an impact on the assessment systems of countries like Norway (Carlsen, 2008), and it is possible that the same influence could be taking place in England where the wide gap between PISA and GCSEs could be narrowing down. One could detect an approximation between PISA and GCSEs in 2007 when the national curriculum went through a reform with changes to GCSE science. Following the 14–19 Green Paper from the government in 2003, the Department for Education and Skills (DfES) asked QCA to revise the key stage 4

[94] Bradshaw et al. (2007), http://www.nfer.ac.uk/nfer/publications/NPC02/NPC02.pdf, p. 15.

[95] Leney, T. and May, T. (2007) *PISA 2006: a first summary of the results for England*, London, QCA. Most observations on PISA and GCSEs exposed in this section are based on this QCA study.

[96] Sturman, L., Ruddock, G., Burge, B., Styles, B., Yin, L. and Vappula, H. (2008) *England's achievement in TIMSS 2007*, DCSF Research Brief DCSF-RBX-18-08, http://www.dcsf.gov.uk/research/data/uploadfiles/TIMSS-2007.pdf (accessed in 23 February 2010)

programme of study for science.[97] The result was a new range of science GCSEs from 2006, with less assessment burden.

Implications for Ofqual

At the end of this study one expects an answer as to how Ofqual can make use of the international studies to fulfil its regulatory remit. From our perspective it is fair to say that the regulator may wish to correlate assessment performance in the national and international assessments because the age of students involved overlap to a great extent:

- PIRLS test 9- to 10-year olds, and the national curriculum assessments for key stage 2 English are aimed at a grade group composed of 10- to 11-year-olds.
- TIMSS mathematics and science test 9- to 10-year-olds (4th grade), whereas key stage 2 national curriculum assessments in mathematics and science are administered to 10- to 11-year-olds.
- PISA surveys involve the 15-year-old age group as they approach the end of compulsory education, and the GCSEs are taken by 14- to 16-year-olds.

Moreover, broad assessment areas in the national and international assessments are comparable with respect to:

- knowledge domains (mathematics and science in NC assessments, TIMSS, PISA, GCSEs)
- basic processes of comprehension in reading assessments (reading in PIRLS and NC assessments)
- cognitive domains such as knowing, applying reasoning (NC assessments and TIMSS)
- some of the competencies involved (NC assessments, PISA and GCSEs).

This study collated evidence to support the view that Ofqual could investigate the adoption of a concurrent validation perspective and ask how well test performance in the national assessments matches expectations based on evidence from the international assessments. A concurrent validation programme would be a meaningful exercise to reassure the public regarding the soundness of national

[97] Association for Science Education (ASE), *Changes to GCSEs in science from 2006 – your questions answered*, http://www.ase.org.uk/htm/homepage/notes_news/january_2005/gcse_faq.pdf (accessed 9 February 2010).

assessments. It is important to note that concurrent validation could be carried out independently from the fact that for the national and international assessments:

- they do not have the same features
- some assessments are high stakes, others are low stakes
- assessments for same age populations show slightly different content specifications
- item formats and amounts of reading involved to understand items may differ
- assessment measurement scales are not the same
- the functions of assessments differ when some assess non-routine problem-solving while others adopt more formal traditional approaches to curricular knowledge
- previous curriculum matching analyses between national and international assessments had demonstrated that national/international curricular match is not a guarantee of success in some of the international assessments.

The last point is particularly intriguing for those who conjectured that the most compelling reason to compare national and international assessments was curriculum overlap. As noted in previous sections, in the event of large overlap between content taught in different countries, pupils' international scores are not necessarily correlated. For example, in 2007 there was a 96 per cent assessment–curriculum match between the national curriculum assessments for key stage 3 mathematics and TIMSS mathematics 8th grade, and the average score for England in the same year was quite low: 513. Yet in 2007 there was only a 70 per cent match between national curriculum assessments for key stage 2 science and TIMSS science 4th grade, when pupils from England scored an average of 542. That is, curriculum match does not necessarily have a direct relationship with good performance in the assessments.

Given the above, there are reasons to support a particular definition of 'comparable assessments' as:

> Two tests involving the same knowledge domains and whose items assess one or more overlapping competencies and cognitive domains, being taken by test takers of the same age groups independently of the size of population tested.

Given the above definition, a concurrent validation exercise with focus on the national curriculum assessments or on GCSEs would treat pupil performance as its object of study or as a *dependent variable* affected by the performance of the same individual in the international assessments. Scores in the international assessments would be

treated as independent variables, whereas performance in the national assessments would be expected to change according to changes detected in scores obtained by the same test takers in the international assessments. Proof of correlation between dependent and independent variables would be an additional tool for regulatory validation studies. The proposed checks would ask how well test takers' performance in the national assessments matches evidence already available from other sources (TIMSS, PIRLS and PISA). Such validation studies would establish the correlation between, for example, performance on cognitive domains (knowing, applying reasoning) or performance on verbal, numerical and abstract components in the national curriculum assessments and GCSEs, and the same *criterion* measures produced by IEA and OECD on a 3- or 4-year basis.

As explained in the Introduction, concurrent validity studies are slightly different, predictive approaches that would ask how well high international assessment performance in TIMSS or on PISA predicts a future criterion, for example high GCSE grades or high performance in GCE A level examinations. These studies would be useful as an extra reference for university selectors. However, it would take a number of years to set up predictive validity. Given the volatile context of regulatory policy making, concurrent validation of the national assessments, and examinations looking at contemporary assessments, could be the most promising to the regulatory remit of safeguarding the validity of the national assessments and examinations.

As argued in this report, PIRLS 2006 and TIMSS 2007 were in many aspects comparable to the national curriculum assessments. Along the same lines, PISA 2006 matched the underlying focuses of assessment in the national curriculum key stage 4, shaping the mathematics and science GCSE papers. TIMSS could be a thermometer for what key stage 2 and 3 pupils are learning and are able to do in mathematics and science. Moreover, now that key stage 3 assessments are no longer statutory in England, TIMSS 8th grade could help in monitoring the achievement of 14-year-old pupils in alternative national low stakes assessments if the key stage 3 curriculum remains closely related to the content assessed by TIMSS.

It would be fair to say that pupil scores in the national curriculum assessments of key stage 2 English could be validated based on their performance in PIRLS. Yet comparisons of reading ability in national and international assessments would be far more complex as one would need to exclude a specific process of comprehension behind national curriculum tests assessment focus 4 on examination and evaluation of content. Against the odds, pupils' performance in national curriculum assessments and PIRLS could still be correlated regarding assessment focuses 2 and 3. As explained, English assessment focus 2 involves pupils' ability to understand, describe, select or retrieve information, events or ideas from texts, and use quotation and reference to text. Assessment focus 3 requires deduction, inference and interpretation of text.

When the Expert Group on Assessment advised the government to adopt national sample assessing at key stage 3 to monitor standards over time, they recommended that: 'where possible, assessment items should be linked to international comparison surveys in which England already participates (for example TIMSS)'.[98] Therefore the Expert Group defended the integration of the international assessments into a more frequent cycle of national sample testing.[99] This Ofqual report supported a slightly different view whereby the international assessments should be used in concurrent validation exercises to address issues that had historically undermined attempts to maintain standards in a weak criterion-referenced system. If it were possible to obtain data on individual test takers, TIMSS could provide evidence of a relationship between individual assessment grades (science and mathematics) and performance on attainment targets that is based on national curriculum holistic level descriptors. These validation studies could prove that candidates with the same 'assessment potential' are rewarded with equivalent levels or grades in national and international assessments. A pupil who scored high on TIMSS would be expected to perform better in the national curriculum assessments than those who received a low score.

Looking at the assessments for 15-year-olds, the PISA mathematics and science surveys, with emphasis on interpretation and application of like skills, assess mostly different cognitive and competency domains, and content domains if compared to the national curriculum assessments and GCSEs. A validity study would need to work around the fact that the percentage of items in key stage 3 mathematics (aimed at 14-year-olds) and PISA (for 15-year-olds) on each of such domains is markedly distinct. Moreover, item formats may also differ in the national and the international assessments, as PISA uses multiple questions to one stem. Correlations between PISA surveys and GCSEs would not be easily defensible. PISA reading literacy surveys may not be particularly useful for concurrent validation of English GCSEs. Ofqual may need to commission further research on PISA 2009 and PISA 2012[100] to follow developments in the OECD survey that had been useful to reveal how good are the like skills of those graduates of England's educational system as opposed to other OECD countries.

---

[98] Bevan, Y., Brighouse, J., Mills, J., Rose, J. and Smith, M. (2009) *Report of the Expert Group on Assessment*, http://publications.dcsf.gov.uk/default.aspx?PageFunction=productdetails&PageMode=publications&ProductId=DCSF-00532-2009 (accessed 9 February 2010).

[99] Ibid., p. 9.

[100] The regulator needs updated studies along the lines of Ruddock et al. (2006).

References

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. European Journal of Psychology of Education, 16, 441-462

Bond, L.A. (1996) Norm- and criterion-referenced testing, *Practical Assessment, Research & Evaluation*, vol. 5, no. 2, http://PAREonline.net/getvn.asp?v=5&n=2 (accessed 9 February 2010).

Bonnet, G. (2002) 'Reflections in a critical eye [1]: on the pitfalls of international assessment. Knowledge and skills for life: first results from PISA 2000', *Assessment in Education*, vol. 9, no. 3, pp. 387–99.

Bradshaw, J., Sturman, L., Vappula, H., Ager, R. and Wheater, R. (2007) *Achievement of 15-year-olds in England: PISA 2006 national report* (OECD Programme for International Student Assessment), Slough, NFER. http://www.nfer.ac.uk/nfer/publications/NPC02/NPC02.pdf

Bramley, T. (2006) 'Equating methods used in KS3 science and English', Paper for the NAA Technical Seminar, Oxford, 23–4 March 2006, Cambridge Assessment.

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O., and Sainsbury, M. (2001). *Framework and Specifications for PIRLS Assessment 2001.* Chestnut Hill, MA: Boston College. http://timssandpirls.bc.edu/pirls2001i/pdf/PIRLS_frame2.pdf

Carlsen, C. (2008) 'The role of testing in an egalitarian society', *Cambridge ESOL Research Notes*, no. 34.

Cronbach, L.J. (1989) 'Construct validation after thirty years', in R.L. Linn (ed.), *Intelligence: measurement, theory, and public policy*, Urbana, IL, University of Illinois Press, pp. 147–71.

DCSF (2007) 'GCSE and equivalent results in England 2006/07', *National Statistics First Release*, SFR 34/2007, http://www.dcsf.gov.uk/rsgateway/DB/SFR/s000754/SFR34-2007_v2.pdf (accessed 9 February 2010).

Downing, S.M. and Haladyna, T.M. (2006) *Handbook of test development*, London, Lawrence Erlsbaum Associates.

Education Reform Act 1988. Office of Public Sector Information http://www.opsi.gov.uk/acts/acts1988/Ukpga_19880040_en_1.htm

Embretson, S. (1983) 'Construct validity: construct representation versus nomothetic span', *Psychological Bulletin*, vol. 93, pp. 179–97.

Feuer, J., Holland, P.W., Bertenthal, M.W., Hemphill, C. and Green, B.F. (1998) *Equivalency and linkage of educational tests: interim report*, Washington, DC, National Academy Press.

Goldstein, H. (2008) 'Comment peut-on utiliser les études comparatives internationale pour doter les politiques educatives d'informations fiables?', *Revue Francaise de Pedagogie*, vol. 164, July/September, pp. 69–76.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., and Brenwald, S. (2008). *Highlights From TIMSS 2007Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2009–001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC., Table 5, p. 10, http://nces.ed.gov/pubs2009/2009001.pdf

Hutchison, D. and Schagen, I. (2006) Comparisons between PISA and TIMSS: are we the man with two watches?, NFER, http://www.brookings.edu/gs/brown/irc2006conference/HutchisonSchagen_paper.pdf (accessed 9 February 2010).

IEA (1996) Beaton, A.; Mullis, I.V.S.; Martin, M.O.; Gonzalez, E.J.; Kelly, D.L.; Smith, T.A. *Mathematics Achievement in the Middle School Years*, Appendix B. http://timss.bc.edu/timss1995i/MathB.html

IEA (2004) Mullis, I.V.S., Martin, M.O., Gonzales, E.J. and Chrostowski, S.J. *TIMSS 2003 international mathematics report*, The Test Curriculum Matching Analysis, Appendix C, http://timss.bc.edu/PDF/t03_download/T03INTLMATRPT.pdf (accessed 22 February 2010),

IEA (2008) Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. *TIMSS 2007International Mathematics report*, Appendix C http://timss.bc.edu/TIMSS2007/PDF/T07_M_IR_AppendixC.pdf (accessed 22 February 2010)

IEA (2007) *TIMSS 2007 technical report*, http://timss.bc.edu/TIMSS2007/PDF/TIMSS2007_TechnicalReport.pdf (accessed 9 February 2010).

Joint Council for Qualifications – JCQ (2008) Entry Trends 2008 – GCSE, Applied GCSE and Entry Level. http://www.jcq.org.uk/attachments/published/1022/GCSE~AppGCSE~Entry%20Trends.pdf

Kane, M.T. (1992) 'An argument-based approach to validity', *Psychological Bulletin*, vol. 112, pp. 527–35.

Massey, A., Green, S., Dexter, T. and Hamnett, L. (2003) *Comparability of national tests over time: KS1, KS2 and KS3 standards between 1996 and 2001*, London, QCA.

Messick, S. (1989) 'Validity', in R.L. Linn (ed.) *Educational measurement* (3rd edn), New York, American Council on Education/Macmillan, pp. 13–103.

Messick, S. (1994) 'The interplay of evidence and consequences in the validation of performance assessments', *Educational Researcher*, vol. 23, no. 2, pp. 13–23.

Messick, S. (1996a) 'Standards-based score interpretation: establishing valid grounds for valid inferences', *Proceedings of the joint conference on standard setting for large-scale assessments*, sponsored by the National Assessment Governing Board and the National Center for Education Statistics, Washington, DC, Government Printing Office.

Messick, S. (1996b) 'Validity of performance assessment', in G. Philips (ed.) *Technical issues in large-scale erformance assessment*, Washington, DC, National Center for Educational Statistics.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O. and Sainsbury, M. (2006) *PIRLS 2006 assessment framework and specifications* (2nd edn), IEA, http://www.ince.mec.es/pub/pirlsmarcosen.pdf (accessed 9 February 2010).

Mullis, I.V.S., Martin, M.O., Kennedy, A.M. and Foy, P. (2007) *PIRLS 2006 international report*, Chestnut Hill, MA, TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O. and Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A. and Galia, J.) (2008a) *TIMSS 2007 international mathematics report: findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*, Chestnut Hill, MA, TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S, Martin, M.O, Ruddock, G.J., O'Sullivan, C.Y., Arora, A. and Erberber, E. (2008b) *TIMSS 2007 assessment frameworks*, TIMSS & PIRLS International Study Center, Boston College, http://timss.bc.edu/TIMSS2007/PDF/T07_AF.pdf (accessed 19 February 2010).

Newton, P.E. (2000) *Maintaining standards over time in national curriculum English and science tests at key stage two. A report for the Qualifications and Curriculum Authority*, NFER,

Newton, P.E. (2005) 'Examination standards and the limits of linking', *Assessment in Education*, vol. 12, no. 2, pp. 105–23.

Newton, P.E. (2008) 'The meaning of maintaining standards. Do we need a new model of awarding? Draft for discussion', Ofqual, 23 October, http://iris/iris/llisapi.dll?func=ll&objId=23674121&objAction=viewversionheader&vernum=1

OECD (2009) *PISA 2006 technical report*, http://www.pisa.oecd.org/dataoecd/0/47/42025182.pdf (accessed 9 February 2010).

Ofqual (2009c) *GCSE subject criteria for English*, September 2001 http://www.ofqual.gov.uk/files/gcse_engcriteria2002.pdf and March 2009 http://www.ofqual.gov.uk/files/2009-03-gcse-eng-subject-criteria.pdf (accessed 19 February 2010)

Olson, J., Martin, M.O., Mullis, I.V.S (eds) (2008) *TIMSS 2007 technical report*, TIMSS & PIRLS International Study Center, Boston College, http://timss.bc.edu/TIMSS2007/PDF/TIMSS2007_TechnicalReport.pdf (accessed 9 February 2010).

QCA (1999) *English programme of study for key stage 2 and attainment targets*, http://curriculum.qcda.gov.uk/key-stages-1-and-2/subjects/english/attainmenttargets/index.aspx?return=/key-stages-1-and-2/subjects/english/keystage2/index.aspx%3Freturn%3D/search/index.aspx%253FfldSiteSearch%253Denglish+Key+stage+2+attainment+targets+%2526btnGoSearch.x%253D39%2526btnGoSearch.y%253D14 (accessed 9 February 2010).

QCA (2006) *Regulatory framework, national curriculum assessments, key stages 1–3*. http://curriculum.qca.org.uk/key-stages-1-and-2/subjects/index.aspx

QCA (2007) *English programme of study for key stage 3 and attainment targets*, http://curriculum.qca.org.uk/uploads/QCA-07-3332-pEnglish3_tcm8-399.pdf?return=/key-stages-3-and-4/subjects/english/keystage3/index.aspx%3Freturn%3D/key-stages-3-and-4/subjects/index.aspx (accessed 9 February 2010).

QCA (2007) *Review of Standards in GCSE English 2002-5.* February 2007, QCA/07/3102. http://www.ofqual.gov.uk/files/QCA-07-3102_standards_GCSE_English_mar07.pdf (acessed 23 February 2010)

QCA (2008) *GCSE qualification criteria*, QCA/07/3165, http://www.qca.org.uk/libraryAssets/media/qca-07-3469_gcse_qualification_criteria.pdf (accessed 9 February 2010).

Robinson, K. (1999) *All our futures: creativity, culture and education*, National Advisory Committee on Creative and Cultural Education.

Ruddock, G., Clausen-May, T., Purple, C. and Ager, R. (2006) *Validation study of the PISA 2000, PISA 2003 and TIMSS 2003 international sStudies of pupil attainment* (DfES Research Report 772), London, DfES.

Ruddock, G. and Sainsbury, M. with Clausen-May, T., Vappula, H., Mason, K., Patterson, E.W., Pyle, K., Kispal, A., Siddiqui, R., McNaughton, S. and Rees, F. (2008) *Comparison of the English core primary curriculum to those of other high performing countries* (DCSF Research Report RW048), London, DCSF, http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/3e/ac/96.pdf

Sainsbury, M. (2003) 'The PIRLS Reading Literacy Tests', Chapter 4, p.32-47 http://www.teachernet.gov.uk/_doc/3980/PIRLS%20full%20report.pdf in Twist, L, Sainsbury, M.; Woodthorpe, A.; Whetton, C. (2003) *Reading all over the world*. NFER Department of Education and Skills. http://nationalstrategies.standards.dcsf.gov.uk/node/88231

Shorrocks-Taylor D. (1999). *National testing: past, present and future* (Issues in Assessment and Testing), Leicester, BPS Books.

Smithers, A. (2004) England's education: what can be learned by comparing countries?', University of Liverpool, http://www.suttontrust.com/reports/pisa_publication.doc

Stiggins, R.J. (1994) *Student-centered classroom assessment*, New York, Merrill.

Sturman, L., Ruddock, G., Burge, B., Styles, B., Yin, L. and Vappula, H. (2008). *England's achievement in TIMSS 2007 (Trends in International Mathematics and Science Study)* (DCSF Research Brief RBX-19-8). London: DCSF. http://www.dcsf.gov.uk/research/data/uploadfiles/TIMSS-2007.pdf (accessed 23 February 2010)

Twist, L., Schagen, I. and Hodgson, C. (2007) *Readers and reading: the national report for England 2006* (PIRLS: Progress in International Reading Literacy Study), Slough, NFER. http://www.nfer.ac.uk/nfer/publications/PRN01/PRN01.pdf

Weir, C.J. (2005) *Language testing and validation*, Basingstoke, Palgrave Macmillan.

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.

Wiliam, D. (2009) 'What do you know when you know the test results? The meanings of educational assessments', *Research Matters*, no. 7, p. 3, http://www.cambridgeassessment.org.uk/ca/digitalAssets/178825_Rearch_Matters_7_2009.pdf (accessed 6 February 2010).

Wise, S. L. & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. Educational Assessment 10, 1-17.

Wolf, L.F., Smith, J.K. and Birnbaum, M.E. (1995) Consequence of Performance, Test Motivation and Mentally Taxing Items. Applied Measurement in Education, Vol. 8, 1995

Ofqual wishes to make its publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2010.

Office of Qualifications and Examinations Regulation
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346

www.ofqual.gov.uk