2

Test Design and Test Development

This chapter describes the test design for PISA 2009 and the processes by which the PISA Consortium, led by ACER, developed the PISA 2009 paper-and-pencil tests for reading, mathematics and science. It also describes the design and development of the computer-based assessment of reading, the digital reading assessment, an innovation in PISA 2009. In the following discussion, the term "reading" generally refers to the core, paper-based reading assessment. The computer-based assessment is referred to as the "digital reading assessment".

TEST SCOPE AND FORMAT

Paper and pencil assessment

In PISA 2009 three subject domains were tested, with reading as the major domain for the second time in a PISA administration and mathematics and science as minor domains.

PISA items are arranged in units based around a common stimulus. Many different types of stimulus are used including passages of text, tables, graphs and diagrams, often in combination. Each unit contains up to five items assessing students' competencies and knowledge.

For the paper-and-pencil assessment there were 37 reading units, comprising a total of 131[1] cognitive items, representing approximately 270 minutes of testing time for reading in PISA 2009. The mathematics assessment consisted of 34[2] items (18 units), a subset of the 48 items used in 2006, representing 90 minutes of testing time. The science assessment consisted of 53 items (18 units), also representing 90 minutes of testing time. The science items were selected from the 108 cognitive items used in 2006.

The 131 cognitive reading items used in the main survey included 26 items from the 2000 test that had been used for linking in 2003 and 2006. A further 11 items from PISA 2000, not used since that administration, were also included. The remaining 94 items were newly developed for PISA 2009. The 11 items retrieved from PISA 2000 and the 94 new items were selected, respectively, from a pool of 24 items retrieved from PISA 2000 and 188 newly-developed items that were tested in a field trial conducted in all countries in 2008, one year prior to the main survey. There was no new item development for mathematics or science.

Item formats employed with reading cognitive items were either selected response multiple choice or constructed response. Multiple-choice items were either standard multiple-choice with four (or in a small number of cases, five) responses from which students were required to select the best answer, or complex multiple-choice presenting several statements for each of which students were required to choose one of several possible responses (yes/no, true/false, correct/incorrect, etc.). Constructed response items were of three broad types. Closed-constructed response items required students to construct a numeric response within very limited constraints, or only required a word or short phrase as the answer. Short response items required a response generated by the student, with a limited range of possible full-credit answers. Open-constructed response items required more extensive writing and frequently required some explanation or justification.

Pencils, erasers, rulers, and in some cases calculators, were provided. It was recommended that calculators be provided in countries where they were routinely used in the classroom. National centres decided whether calculators should be provided for their students on the basis of standard national practice. No test items required a calculator, but some mathematics items involved solution steps for which the use of a calculator could be of assistance to some students.

Digital Reading Assessment (DRA)

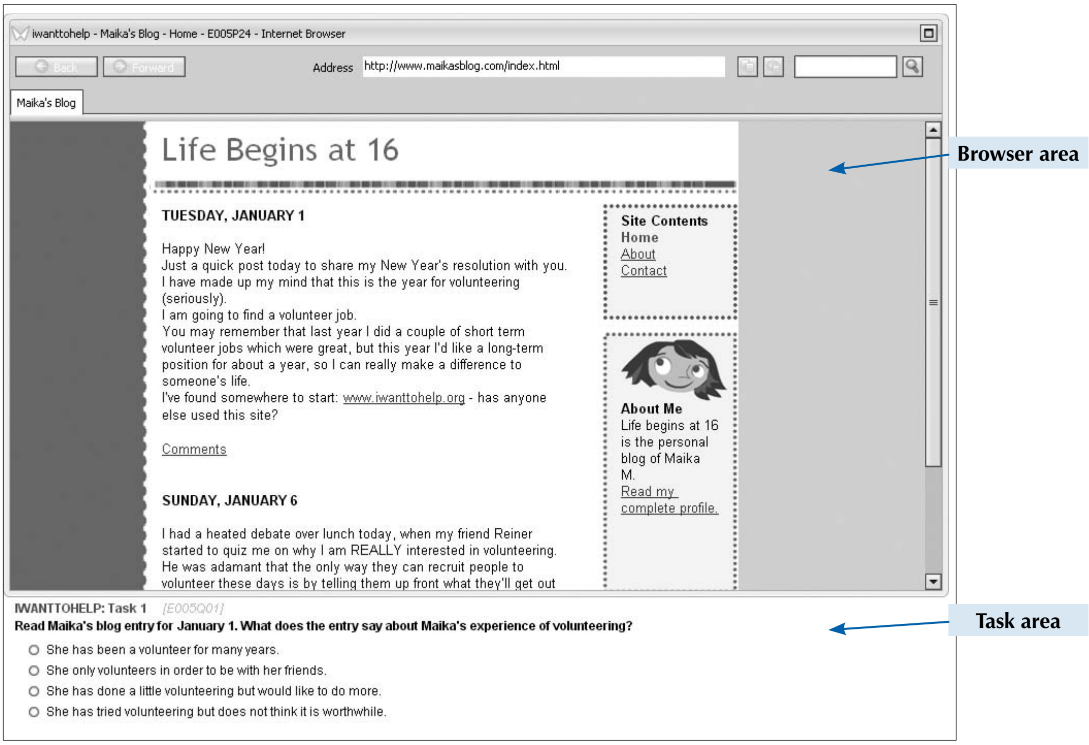
For PISA 2009, countries were offered an assessment of reading in a digital environment (DRA), as an international option.

As with the paper-and-pencil assessment of reading, digital reading items are arranged in units based around a common stimulus, but the stimulus used in the digital reading assessment comprises digital texts with the structures and features of websites, e-mails, blogs and so on. Each unit contains up to four items assessing students' competencies and knowledge.

The digital reading assessment comprised nine units, with a total of 29 items, representing approximately 60 minutes of testing time. These items were selected from a pool of 72 newly-developed digital reading items that were tested in a field trial conducted in all countries participating in the international option in 2008, one year prior to the main survey.

In the digital reading assessment, the screen has two areas: a browser area, in which the stimulus is displayed, and a task area, in which the questions are provided. Figure 2.1 shows the screen layout.

■ Figure 2.1 ■
Screen layout for the Digital Reading Assessment



iwanttohelp - Maika's Blog - Home - E005P24 - Internet Browser

Address http://www.maikasblog.com/index.html

Maika's Blog

Life Begins at 16

TUESDAY, JANUARY 1

Happy New Year!
Just a quick post today to share my New Year's resolution with you.
I have made up my mind that this is the year for volunteering (seriously).
I am going to find a volunteer job.
You may remember that last year I did a couple of short term volunteer jobs which were great, but this year I'd like a long-term position for about a year, so I can really make a difference to someone's life.
I've found somewhere to start: www.iwanttohelp.org - has anyone else used this site?

Comments

SUNDAY, JANUARY 6

I had a heated debate over lunch today, when my friend Reiner started to quiz me on why I am REALLY interested in volunteering.
He was adamant that the only way they can recruit people to volunteer these days is by telling them up front what they'll get out

Site Contents
Home
About
Contact

About Me
Life begins at 16 is the personal blog of Maika M.
Read my complete profile.

IWANTTOHELP: Task 1 {E005Q01}
Read Maika's blog entry for January 1. What does the entry say about Maika's experience of volunteering?

○ She has been a volunteer for many years.
○ She only volunteers in order to be with her friends.
○ She has done a little volunteering but would like to do more.
○ She has tried volunteering but does not think it is worthwhile.

For most items, students provided their responses in the task area. Item formats employed were selected response or constructed response. Most of the selected-response items were in multiple-choice format of the standard type, in which students are required to select the best answer from a set of four options in the task area. A variation on multiple-choice, exploiting the interactive possibilities of the medium, involves students selecting an option from a dropdown menu in the browser area. Open-constructed response items require more extensive writing and frequently require some explanation or justification. Responses were given either in a text box in the task area, or, where appropriate, in the browser area in the form of an e-mail message.

TEST DESIGN

Paper-based assessment

The standard main survey items were allocated to thirteen item clusters (seven reading clusters, three mathematics clusters and three science clusters) with each cluster representing 30 minutes of test time. The items were presented to students in thirteen standard test booklets, with each booklet being composed of four clusters. R1 to R7 denote the reading clusters, M1 to M3 denote the mathematics clusters, and S1 to S3 denote the science clusters. R1 and R2 were the same two reading clusters as those administered in 2003 and 2006. The mathematics clusters were three of the four intact clusters used in 2006 (M1 from 2006 was omitted). The three science clusters were not intact clusters from PISA 2006; items were selected from across the 2006 main survey pool to represent that pool as closely as possible in terms of competency and knowledge classifications, item format types, range of difficulty, layout and cluster position.

In addition to the thirteen two-hour booklets, a special one-hour booklet, referred to as the UH Booklet (Une Heure booklet), was prepared for use in schools catering for students with special needs. The UH Booklet contained about half as many items as the other booklets, with about 50% of the items being reading items, 25% mathematics and 25% science. The items were selected from the main survey items taking into account their suitability for students with special educational needs.

The cluster rotation design for the standard main survey is shown in Table 2.1.

Table 2.1 **Cluster rotation design used to form standard test booklets for PISA 2009**

Booklet ID	Cluster			
1	M1	R1	R3A	M3
2	R1	S1	R4A	R7
3	S1	R3A	M2	S3
4	R3A	R4A	S2	R2
5	R4A	M2	R5	M1
6	R5	R6	R7	R3A
7	R6	M3	S3	R4A
8	R2	M1	S1	R6
9	M2	S2	R6	R1
10	S2	R5	M3	S1
11	M3	R7	R2	M2
12	R7	S3	M1	S2
13	S3	R2	R1	R5
UH	**Reading**	**Mathematics** **Science**		

The fully-linked design is a balanced incomplete block design. Each cluster appears in each of the four possible positions within a booklet once and so each test item appears in four of the test booklets. Another feature of the design is that each pair of clusters appears in one (and only one) booklet.

Each sampled student was randomly assigned one of the thirteen booklets administered in each country, which meant each student undertook two hours of testing. Students were allowed a short break after one hour.

In PISA 2009 some countries were offered the option of administering an easier set of booklets. The offer was made to countries that had achieved a mean scale score in reading of 450 or less in PISA 2006, and to new countries that were expected – judging by their results on the PISA 2009 field trial conducted in 2008 – to gain a mean result at a similar level. The purpose of this strategy was to obtain better descriptive information about what students at the lower end of the ability spectrum know, understand and can do as readers. A further reason for including easier items was to make the experience of the test more satisfying for individual students with very low levels of reading proficiency. For countries that selected the easier set of booklets two of the standard reading clusters (R3A and R4A) were substituted with two easier reading clusters (R3B and R4B). Apart from level of difficulty, the sets of items in the standard and easier clusters were matched, in terms of the distribution of text format, aspect and item format. The other eleven clusters (five clusters of reading items, three clusters of mathematics items and three clusters of science items) were administered in all countries.

Table 2.2 shows the full test design used in the 2009 main survey.

Table 2.2 **Cluster rotation design used to form all test booklets for PISA 2009**

Booklet ID	Cluster				Standard booklet set	Easier booklet set
1	M1	R1	R3A	M3	Y	
2	R1	S1	R4A	R7	Y	
3	S1	R3A	M2	S3	Y	
4	R3A	R4A	S2	R2	Y	
5	R4A	M2	R5	M1	Y	
6	R5	R6	R7	R3A	Y	
7	R6	M3	S3	R4A	Y	
8	R2	M1	S1	R6	Y	Y
9	M2	S2	R6	R1	Y	Y
10	S2	R5	M3	S1	Y	Y
11	M3	R7	R2	M2	Y	Y
12	R7	S3	M1	S2	Y	Y
13	S3	R2	R1	R5	Y	Y
21	M1	R1	R3B	M3		Y
22	R1	S1	R4B	R7		Y
23	S1	R3B	M2	S3		Y
24	R3B	R4B	S2	R2		Y
25	R4B	M2	R5	M1		Y
26	R5	R6	R7	R3B		Y
27	R6	M3	S3	R4B		Y
UH	**Reading**	**Mathematics** **Science**				

Although only two of the clusters differed for standard and easier administration, the cluster rotation in the booklets (where each cluster appears four times) means that more than half of the booklets are affected by the alternatives. Countries administering the standard set of booklets implemented Booklets 1 to 13. Countries administering the easier set of booklets implemented Booklets 8 to 13 and Booklets 21 to 27.

Digital Reading Assessment

The main survey items for the digital reading assessment were allocated to three item clusters with each cluster representing 20 minutes of test time. The items were presented to students in six test forms, with each form being composed of two clusters according to the rotation design shown in Table 2.3.

Table 2.3 **Digital reading assessment test design**

	Cluster 1	Cluster 2
Test 1	A	B
Test 2	B	A
Test 3	B	C
Test 4	C	B
Test 5	C	A
Test 6	A	C

Each cluster is paired with each of the other clusters in two forms, once in the first position and once in the second position. Each sampled student was randomly assigned one of the six forms, which meant each student undertook 40 minutes of testing.

Each unit consisted of several items referring to a common stimulus, comprising multiple linked browser pages. Following the advice of the DRA Advisory Group, units and items within units were delivered in a fixed order, or lockstep fashion. This meant that students were not able to return to an item or unit once they had moved to the next item/unit. Each time a student clicked the 'Next' test navigation button, a dialog box displayed a warning that the student was about to move on to the next item and that it would not be possible to return to previous items. At this point students could either confirm that they wanted to move on or cancel the action and return to the item they had been viewing.

Lockstep delivery enabled test developers to specify the starting browser page for each item. This meant that all students began in the same place within the stimulus and, if they had previously navigated through a series of less relevant pages, did not have to spend time finding their way to an appropriate page to begin the item task.

TEST DEVELOPMENT CENTRES

Experience gained in the three previous PISA assessments showed the importance of using the development expertise of a diverse range of test centres to help achieve conceptually rigorous material that has the highest possible levels of cross-cultural and cross-national diversity. Accordingly, to prepare new reading items for PISA 2009 the Consortium drew on the resources of five test development centres in culturally-diverse and well-known institutions, namely ACER (Australia), aSPe (University of Liege, Belgium), ILS (University of Oslo, Norway), DIPF (Germany) and NIER (Japan) (see Annex H).

In addition, for PISA 2009 the test development teams were encouraged to conduct initial development of items, including cognitive laboratory activities, in their local language. Translation to the OECD source languages (English and French; English only for the digital reading assessment) took place only after items had reached a well-formed state. The work of the test development teams was coordinated and monitored overall at ACER by the Consortium's manager of test and framework development for reading.

DEVELOPMENT TIMELINE

The PISA 2009 project started formally in August 2006, and concluded in December 2010. Planning for item development began in June 2006, with preparation of material for a two-day meeting of test developers from each test development centre, which was held in Frankfurt on 30-31 August, 2006. The meeting had the following purposes:

- to become familiar with the PISA 2000 reading literacy framework, especially its implications for test development;
- to discuss the requirements for item development, including item presentation and formats, use of templates and styles and cognitive laboratory procedures and timelines;

- to discuss factors that influence item difficulty, particularly in light of the intention to develop items at the extremes of the scale (a contractual requirement);
- to be briefed on detailed guidelines, based on experience from the first three PISA administrations, for avoiding potential translation and cultural problems when developing items; and
- to review sample items prepared for the meeting by each of the test development centres.

Test development began in earnest after the first PISA 2009 Reading Expert Group (REG) meeting which was held in Lyon on 5–7 October 2006. The main phase of test development finished when the items were distributed for the field trial in December 2007. During this 15-month period, intensive work was carried out writing and reviewing items, and on various cognitive laboratory activities. The field trial for most countries took place between March and August 2008, after which items were selected for the main survey and distributed to countries in December 2008.

Table 2.4 shows the major milestones and activities of the PISA 2009 test development timeline.

Table 2.4 Test development timeline for PISA 2009

Activity	Period
Review of 2000 reading framework and development of 2009 reading framework	October 2006 – February 2009
First phase item development in English (paper-based and computer-based) and French (paper-based)	June 2006 – October 2007
Item development workshop for participating countries	March 2007
Item submissions from countries	February – June 2007
Distribution of field trial material	November – December 2007
Translation into national languages	November 2007 – April 2008
Field trial coder training	February 2008
Field trial in participating countries	March – September 2008
Selection of items for main survey	August – October 2008
Preparation of final source versions of all main survey materials, in English (paper-based and computer-based) and French (paper-based)	October – December 2008
Distribution of main survey material	November – December 2008
Main survey coder training	February 2009
Main survey in participating countries	March – September 2009

THE PISA 2009 READING LITERACY FRAMEWORK

For each PISA subject domain, an assessment framework is produced to guide the PISA assessments in accordance with the policy requirements of the OECD's PISA Governing Board (PGB). The framework defines the domain, describes the scope of the assessment, specifies the structure of the test – including item format and the preferred distribution of items according to important framework variables – and outlines the possibilities for reporting results.

The PISA domain frameworks are conceived as evolving documents that will be adapted over time to integrate developments in theory and practice. Since a framework for PISA reading had been developed for the first PISA administration in 2000, the PISA 2009 work began with a review of the existing framework at the initial REG meeting in October 2006. It was agreed that much of the substance of the PISA 2000 framework should be retained for PISA 2009, but new elements were to be added or given additional emphasis: notably, the incorporation of digital reading, and the elaboration of engagement and metacognition in reading (subsequently called "reading strategies"). Re-drafting of the framework commenced in the ensuing months, guided by the REG under the leadership of its Chair, Irwin Kirsch.

The OECD invited national experts to a reading forum held in February 2007, to review the first draft of a revised and expanded reading framework for PISA 2009. A further draft was then produced, and considered by the PGB at its meeting in Oslo in March 2007. After the PGB meeting further revisions were made, culminating in the submission of a new draft to the PGB in July 2007. This version substantially remained unchanged and guided test development and selection for both print reading and DRA for the field trial and the main survey.

In early 2009 the framework was prepared for publication along with an extensive set of example items. All three PISA 2009 cognitive frameworks (as well as the questionnaire framework) were published in *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science* (OECD, 2010a). The mathematics and science frameworks were unchanged from 2006.

ITEM DEVELOPMENT PROCESS

The item development process commenced with preparations for the meeting of test developers held in Frankfurt in August 2006. This included the preparation of documentation to guide all parts of the process for the development of cognitive items. The process continued with the calling of submissions from participating countries, writing and reviewing items, carrying out pilot tests of items and conducting an extensive field trial, producing final source versions of all items in both English and French (for digital reading, in English only), preparing coding guides and coder training material, and selecting and preparing items for the main survey.

Since a similar process was followed for the development of print and digital reading items, it should be assumed that the following description applies to both, except where a variation is explicitly stated.

Cognitive item development was guided by a set of documents prepared iteratively over preceding administrations of PISA, augmented by discussion at the test development meeting. The orientation included an overview of the development process and timelines, a specification of item requirements, including the importance of framework fit, and a discussion of issues affecting item difficulty. These principles were expected to be followed by item developers at each of the five test development centres. They were later incorporated into the document *Item Submission Guidelines for Reading – PISA 2009*.[3]

A complete PISA unit consists of some stimulus material, one or more items (questions), and a guide to the coding of responses to each question. Each coding guide comprises a list of response categories (full, partial and no credit), each with its own scoring code, descriptions of the kinds of responses to be assigned each code, and sample responses for each response category.

First phase of development

Typically, the following steps were taken in the first phase of the development of reading items originating at a test development centre. The steps are described in a linear fashion, but in reality they were often negotiated in a cyclical fashion, with items going through the various steps more than once.

Initial preparation

Selection of stimulus is a key component of reading test development. In the case of print reading material, test developers in each of the five Consortium test development centres found potential stimulus and exchanged it with other centres (in English translation if necessary) to ascertain whether colleagues agreed that it was worth developing further. The stimulus was formatted even at this early stage in a manner similar to that planned for the final presentation. In the case of digital reading, three of the Consortium test development centres – ACER, aSPe and DIPF – developed digital stimulus: screen-shot mock-ups of stimulus pages were created, with accompanying descriptions of the navigation features envisaged for each page.

For those pieces of stimulus that were judged worth pursuing, test developers prepared units in both English and their native language in a standard format, including stimulus, several items (questions), and a proposed coding guide for each item. Items were then subjected to a series of cognitive laboratory activities: item panelling (also known as item shredding or cognitive walkthrough), cognitive interviews, and pilot or pre-trial testing (also known as cognitive comparison studies).

Local item panelling

Each unit first underwent extensive scrutiny at a meeting of members of the originating test development team. This stage of the cognitive laboratory process typically involved item writers in a vigorous analysis of all aspects of the items from the point of view of a student, and from the point of view of a coder.

Items were revised, often extensively, following item panelling. When substantial revisions were required, items went back to the panelling stage for further consideration.

Cognitive interviews

Many units were then prepared for individual students or small groups of students to attempt. For print reading material a combination of think-aloud methods, individual interviews and group interviews was used with students to ascertain the thought processes typically employed as students attempted the items. For digital reading items, all cognitive interviews were conducted individually, using either audio-recording of responses and screen capture, or dual administration, with one researcher interacting with the student and another researcher observing and recording navigation behaviour.

Items were revised, often extensively, following their use with individuals and small groups of students. This stage was particularly useful in clarifying the wording of questions, and gave information on likely student responses that was used in refining the response coding guides.

Local pilot testing

As the final step in the first phase of print item development, sets of units were piloted with several classes of 15-year-olds. As well as providing statistical data on item functioning, including the relative difficulty of items, this enabled real student responses derived under formal test conditions to be obtained, thereby enabling more detailed development of coding guides.

Pilot test data were used to inform further revision of items where necessary or sometimes to discard items altogether. Units that survived relatively unscathed were then formally submitted to the test development manager to undergo their second phase of development.

Second phase of development

The second phase of item development began with the review of each unit by at least one test development team that was not responsible for its initial development. Each unit was then included in at least one of a series of pilot studies with a substantial number of students of the appropriate age.

International item panelling

The feedback provided following the scrutiny of items by international colleagues often resulted in further improvements to the items. Of particular importance was feedback relating to the operation of items in different cultures and national contexts, which sometimes led to items or even units being discarded. Surviving units were considered ready for further pilot testing and for circulation to national centres for review.

International pilot testing

For each pilot study, test booklets were formed from a number of units developed at different test development centres. These booklets were trial tested with several whole classes of students in several different schools. Field-testing of this kind mainly took place in schools in Australia because of translation and timeline constraints. Sometimes, multiple versions of items were trialled and the results were compared to ensure that the best alternative form was identified. Data from the pilot studies were analysed using standard item response techniques. For digital reading items, international pilot testing was not possible due to technical constraints at this stage of development. However some cognitive interviews with individual students were conducted in school settings.

Many items were revised, usually in a minor fashion, following review of the results of pilot testing. If extensive revision was considered necessary, the item was either discarded or the revised version was again subject to panelling and piloting. One of the most important outputs of this pilot testing was the generation of many student responses to each constructed-response item. A selection of these responses was added to the coding guide for the item to further illustrate each response category and provide more guidance for coders.

National item submissions

An international comparative study should ideally draw items from as many participating countries as possible to ensure wide cultural and contextual diversity. A comprehensive set of guidelines, was developed to encourage and assist national submission of reading items. The document *Item Submission Guidelines for Reading – PISA 2009* was distributed to PISA 2009 National Project Managers (NPMs) in February 2007.

The guidelines described the scope of the item development task for PISA 2009, the arrangements for national submissions of items and the item development timeline. In addition, the guidelines contained a detailed discussion of item requirements and an overview of the full item development process for PISA 2009.

To assist countries in submitting high quality and appropriate material, ACER conducted a one-day reading item development workshop for interested national centres at the end of the first NPM meeting for PISA 2009, in March 2007. It was attended by 30 individuals from 22 national centres.

The due date for national submission of items was 29 June 2007, as late as possible given field trial preparation deadlines. Items could be submitted in English, French, German, Spanish, Japanese or Italian. Countries were urged to submit items as they were developed, rather than waiting until close to the submission deadline. It was emphasised that before items

were submitted they should have been subject to some cognitive laboratory activities involving students, and revised accordingly. An item submission form was provided with the guidelines and a copy had to be completed for each unit, indicating the source of the material, any copyright issues, and the framework classifications of each item.

For print reading, a total of 162 units were processed from 30 countries. Countries submitting units were: Argentina, Belgium, Brazil, Canada, Chile, Colombia, the Czech Republic, Denmark, Finland, France, Greece, Hungary, Ireland, Korea, Lithuania, Macao-China, Mexico, the Netherlands, New Zealand, Norway, Portugal, Qatar, Serbia, the Slovak Republic, Spain, Sweden, Switzerland, the United Kingdom, the United States and Uruguay. Most countries chose to submit their material in English, but submissions were also received in French, German and Spanish.

For the digital reading assessment, seven units were submitted by Canada. Five of these units were submitted in English and two in French. In addition, one unit submitted by Lithuania as a print reading unit was judged by the REG to be more suitable as a digital reading unit.

Some submitted units had already undergone significant development work, including pilot testing, prior to submission. Others were in a less developed state.

For print reading, all of the units submitted were reviewed by at least two of the test development centres, apart from a small number (about 10%) where ACER judged that the material too closely duplicated something that had already been developed for the 2009 pool, or was part of the trend or previously released material. Less than 30% of the units were deemed unsuitable for the PISA 2009 reading assessment in the review by two test development centres. Reasons for assessing units as unsuitable included inappropriate content (e.g. material that might be considered offensive in some countries), cultural bias and ephemerality.

The remaining units, in excess of 60% of those submitted, were considered suitable, though not all were able to be used. Various criteria were used to select those that were actually used, including overall quality of the unit, amount of revision required, and framework coverage. Consistent with the advice provided to countries, early submissions had a greater chance of selection than those received towards the end of the submission period. Nevertheless, high importance was placed on including units from as wide a range of countries as possible and, as a result, only six of the submitting countries “missed out”. Some quite good units were not included solely because their content overlapped too much with at least one existing unit.

Since only one national centre submitted material for DRA, the review process was informal, with the unit selected for development discussed in detail with the submitting country (Canada).

For print reading, units requiring further initial development were distributed among the test development centres. Typically, after local panelling and revision, they were fast-tracked into the second phase of item development as there was rarely time for cognitive interviews or pilot testing to be conducted locally. However, all these units underwent international pilot testing (as described above), along with the units that originated at test development centres.

A total of 31 print reading units and two digital reading units from national submissions were included in the bundles of items (four print reading, and four digital reading) circulated to national centres for review. Feedback was provided to countries on any submitted units that were not used. This practice, together with the provision of an item development workshop for national centre representatives early in a cycle, should contribute to improvements in the quality of national submissions in the future.

National review of items

In February 2007, NPMs were given a set of item review guidelines to assist them in reviewing cognitive items and providing feedback. At the same time, NPMs were given a schedule for the distribution and review of bundles of draft items during the remainder of 2007. A central feature of those reviews was the requirement for national experts to rate items according to various aspects of their relevance to 15-year-olds, including whether they related to material included in the country’s curriculum, their relevance in preparing students for life, how interesting they would appear to students and their authenticity as real applications of reading. NPMs were also asked to identify any cultural concerns or other problems with the items, such as likely translation or coding difficulties, and to give each item an overall rating for retention in the item pool.

As items were developed to a sufficiently complete stage, they were despatched to national centres for review. Four bundles of print reading items were distributed. The first bundle, including 8 units (52 items) was despatched on 14 February 2007. National centres were provided with an Excel® worksheet, already populated with unit names and item identification codes, in which to enter their ratings and other comments. Subsequent bundles were despatched on 16 April (17 units, 133 items), 16 July (18 units, 117 items) and 9 August (19 units, 124 items). In general, except for the last bundle, about four weeks was allowed for feedback.

For DRA, four bundles of items were distributed. The first bundle, including 5 units (35 items) was released on 30 April 2007. Subsequent bundles were despatched on 3 September (8 units, 58 items), 12 October (4 units, 35 items) and 19 October (4 units, 35 items). For digital reading, an online item review system was established, allowing countries to view the stimulus and items in digital format and to enter their ratings and comments on the material in a computer-based questionnaire format. The criteria for rating the material were similar in substance to those called for in the print reading review, with the addition of a question about the technological demands of the assessment items.

For each bundle, a series of reports was generated summarising the feedback from NPMs. The feedback frequently resulted in further revision of the items. In particular, cultural issues related to the potential operation of items in different national contexts were highlighted and sometimes, as a result of this, items had to be discarded. Summaries of the ratings assigned to each item by the NPMs were used extensively in the selection of items for the field trial.

International item review

As well as the formal, structured process for national review of items, cognitive items were also considered in detail, as they were developed, at meetings of the REG that took place in October 2006 and February, June and September 2007. The REG members were also invited to submit comments and ratings of the items as they were released in bundles.

Reading for School questionnaire

It was proposed to include a short questionnaire, Reading for School, at the end of the cognitive booklets. The focus of the questionnaire was to be on school-based reading, whether done in the classroom or for homework, with the purpose of collecting information about reading curriculum and pedagogy as experienced by 15-year-olds. The questions were developed from surveys administered in previous international studies (Grisay, 2008; Purves, 1973), which had investigated school-aged students' opportunities to read different materials, and the ways in which reading was taught. For the PISA Reading for School questionnaire, items were designed to align with the PISA reading framework, so that links could potentially be made between reading practices at school and the proficiency of students in various parts of the PISA reading assessment. Consequently, questions were developed that asked about the kinds of texts (based on the text formats and text types defined in the framework) and the kinds of reading tasks (aligned with the aspects of reading) that 15-year-olds encountered in school-based reading.

The items underwent an extensive series of reviews by researchers and test developers, and were submitted to cognitive laboratory procedures (item panelling and cognitive interviews) in Australia, Finland, Japan, the Netherlands and Norway. Three sets of Reading for School items (Forms A, B and C), each designed to take about five minutes to complete, were assembled for the field trial.

Preparation of dual (English and French) source versions

Both English and French source versions of all paper-based test instruments were developed and distributed to countries as a basis for local adaptation and translation into national versions. An item-tracking database, with web interface, was used by both test developers and Consortium translators to access items. This ensured accurate tracking of the English language versions and the parallel tracking of French translation versions.

Part of the translation process involved a technical review by French subject experts, who were able to identify issues with the English source version related to content and expression that needed to be addressed immediately, and that might be of significance later when items would be translated into other languages. Many revisions were made to items as a result of the translation and technical review process, affecting both the English and French source versions. This parallel development of the two source versions assisted in ensuring that items were as culturally neutral as possible, identified instances of wording that could be modified to simplify translation into other languages, and indicated where additional translation notes were needed to ensure the required accuracy in translating items to other languages.

For DRA, only an English source version was developed.

FIELD TRIAL

The PISA field trial was carried out in most countries in the first half of 2008. An average of over 200 student responses to each item was collected in each country. During the field trial, the Consortium set up a coder query service. Countries were encouraged to send queries to the service so that a common adjudication process was consistently applied to all coders' questions about constructed-response items. Between July and November 2008, the test development centres, the REG and national centres reviewed the field trial data to recommend a selection of field trial items for the main survey.

Field trial selection

Print reading

A total of 62 reading units (425 cognitive items) were circulated to national centres for review from February to August 2007. After consideration of country feedback, 53 units (348 cognitive items) were retained as the pool of units to be considered by the REG for inclusion in the field trial. Twenty-seven of these units (51% of the items) originated in national submissions.

The cognitive items to be used in the 2008 field trial were selected from the item pool at the meeting of the REG held in Dubrovnik in mid-September 2007. The selection process took two days to complete.

At the beginning of the process, REG members were provided with a report on the final pool of reading items available for selection for the field trial. The report included a summary of the item development process for PISA 2009 and detailed item reports, including the classification of all items according to their Framework characteristics, estimates of difficulty and average ratings given by NPMs.

For the purposes of item selection, the units were divided into three groups: non-continuous and mixed texts, continuous texts and easy units. For each of these three groups of units, REG members worked, first in small groups, then in plenary, to nominate a set of units for inclusion in the field trial. The discussion was based on the selection criteria outlined for the field trial items, as well as the report on the final pool of items. REG members were not aware of the origin of any of the material.

Having made the selection of units for inclusion in the field trial, the REG then selected individual items from within the chosen units. In order to inform the discussion on item choice, a report on factors influencing item difficulty was presented. The REG members then made their item selection in the same way as their unit selection, working first in groups, then in plenary to select items from non-continuous and mixed texts, then continuous texts and, finally, easy units.

The characteristics of the selected items, including framework classifications and estimated difficulties, were then examined. Minor adjustments were made to match framework requirements. This revised selection was approved by the REG (allowing for further minor adjustments to be made by test developers), and subject to NPM and PGB endorsement.

The REG recommended selection for print reading was presented to a meeting of NPMs in the week after the REG meeting. The NPMs endorsed the REG's recommended selection.

Subsequently a small number of items had to be dropped because of space and layout constraints when the Consortium test developers assembled the units into clusters and booklets. The final field trial item pool for print reading included a total of 240 reading items, comprising 24 items retrieved from PISA 2000, 28 link items (items that had been administered in every cycle since 2000 to collect trend data) and 188 new items.

Table 2.5 **Print reading field trial cognitive items**

New items	188
PISA 2000 retrieved items	24
Link items	28
Total	**240**

Digital reading

For digital reading, from April to October four bundles of items were released for online review. While national centres were invited to review 21 units (163 items) during this phase, all items reviewed in the first bundle were revised and then re-released in subsequent bundles, so that only 16 units (128 individual items) in total were in the pool for field trial selection. The later development cycle of digital reading items meant that REG and NPM meetings in September 2007 did not have the full set of items available for selection. Consequently, REG and NPM feedback, via the online review system, was used by the Consortium to inform the selection of digital reading items for the field trial. Eighteen participants provided feedback and this was generally very favourable. Table 2.6 summarises quantitative responses on a scale from 1 (lowest) to 5 (highest) and gives comparative information for the print assessment items.

Table 2.6 **Average ratings for DRA tasks from national centres**

	Relevance to school	Relevance to life beyond school	Interest level	Priority for inclusion
DRA tasks	3.95	4.25	3.93	3.95

After consideration of country feedback, 13 units, comprising a total of 72 tasks, were selected for inclusion in the field trial. In addition, a practice test comprising two units (10 tasks) and an effort thermometer task were produced.

The practice test was designed to familiarise students with the DRA interface. It described the layout of the screen, the methods of navigation that were possible, explained how to keep track of the time left for their testing session and how their progress throughout the test was displayed, and provided exercises on how to use the stimulus elements (such as links, tabs, drop-down menus) and respond to questions in the computer based environment (e.g. through text input or selection of radio-buttons).

The effort thermometer task was administered at the end of the digital reading assessment. The purpose of the task, which was modelled on the an effort thermometer instrument administered at the end of the paper-based cognitive booklets in PISA 2003 and PISA 2006, was to collect information about students' motivation when completing the digital reading assessment. Students were asked to indicate the amount of effort they put into doing the digital reading assessment compared with a school test, and compared with the paper-based PISA assessment that they had recently completed. However, after examining the results it became clear that many students did not interpret "effort" in a motivational sense when comparing the digital and paper-based assessments (the digital reading assessment was much shorter and therefore required less effort). So the effort thermometer was not carried forward into the main survey of the digital reading assessment.

Field trial design

Paper-based assessment

The field trial design for the paper-based assessment comprised 16 clusters of reading items (R1 to R16), 3 clusters of mathematics items (M1 to M3) and 3 clusters of science items (S1 to S3).

Clusters R1 and R2 were intact clusters that had been used in PISA 2003 and 2006, comprising 8 link units (28 items). The 35 new reading units and 6 units retrieved from PISA 2000 were allocated to 14 clusters, R3 to R16.

M1, M2 and M3 were 3 intact mathematics clusters from PISA 2006 comprising 35 items (18 units). S1, S2 and S3 were 3 science clusters comprising 53 items (18 units) selected from the 108 cognitive items used in 2006.

Nine regular two-hour booklets, each comprising four clusters, were administered in the field trial. Each cluster was designed to take up 30 minutes of testing time, thus making up booklets with two hours' worth of testing time. New reading clusters appeared once in the first half of a booklet and once in the second half, in booklets 1 to 7, and were administered in all participating countries. The reading, mathematics and science link material appeared in booklets 8 and 9; these booklets were administered only in countries participating in PISA for the first time in 2009. All nine regular booklets included one of three sets of Reading for School items (Form A, B or C). This short questionnaire was administered immediately following the cognitive assessment and was designed to take about five minutes to complete.

In addition, the field trial design included a one-hour test booklet of two of the new reading clusters, R3 and R4, for special educational needs students. Items in these clusters were selected taking into account their suitability for students with special educational needs.

Table 2.7 shows the field trial design for the paper-based assessment.

Table 2.7 **Allocation of item clusters to test booklets for field trial**

Booklet	Cluster				Reading for School survey
1	R3	R10	R12	R4	A
2	R4	R11	R13	R5	B
3	R5	R12	R14	R6	C
4	R6	R13	R15	R7	A
5	R7	R14	R16	R8	B
6	R8	R15	R10	R9	C
7	R9	R16	R11	R3	A
8	R1	M1	M2	M3	B
9	R2	S1	S2	S3	C
UH	**R3**	**R4**			

Digital reading assessment

The 13 field trial units were arranged into five 20-minute clusters to allow the construction of five 40-minute test forms. Each cluster appeared first in one test form and second in another form (AB, BC, CD, DE, EA)

Despatch of field trial instruments

Final English and French paper-based source versions of field trial units were distributed to national centres in two despatches, on 12 October (link units) and 30 November (new reading units). Clusters and booklets were distributed on 17 December 2007 in both Microsoft Word® and PDF formats. All material could also be downloaded from the PISA website from the time of despatch.

Revised versions of the digital reading items, accompanied by their coding guides, were released for adaptation and translation in the period late November to early December 2007.

National centres then commenced the process of preparing national versions of all units, clusters and booklets. All items went through an extremely rigorous process of adaptation, translation and external verification in each country to ensure that the final test forms used were equivalent. That process and its outcomes are described in Chapter 5.

Field trial coder training

Following final selection and despatch of items to be included in the field trial, various documents and materials were prepared to assist in the training of response coders. International coder training sessions for reading, mathematics and science were scheduled for 25–29 February 2008. For the paper-based assessments, consolidated coding guides were prepared, in both English and French, containing all those items that required manual coding. The guides emphasised that coders were to code rather than score responses. That is, the guides separated different kinds of possible responses, which did not all necessarily receive different scores. A separate training workshop document in English only was also produced for each paper-based domain. These workshop documents contained additional student responses to the items that required manual coding, and were used for practice coding and discussion at the coder training sessions. For digital reading, a combined coding guide and workshop document was produced in English only.

Countries sent representatives to the training sessions, which were conducted in Offenbach, Germany. Open discussion of how the workshop examples should be coded was encouraged and showed the need to introduce a small number of amendments to coding guides. These amendments were incorporated in a final despatch of coding guides and training materials on 6 March 2008. Following the international training sessions, national centres conducted their own coder training activities using their verified translations of the consolidated coding guides. The support materials for coding prepared by the Consortium included a coder recruitment kit to assist national centres in recruiting people with suitable qualifications as expert coders.

Field trial coder queries

The Consortium provided a coder query service to support the coding of constructed-response items in each country. When there was any uncertainty, national centres were able to submit queries by e-mail to the query service, and they were immediately directed to the relevant Consortium expert. Considered responses were quickly prepared, ensuring greater consistency in the coding of responses to items.

The queries with the Consortium's responses were published on the PISA website. The queries report was regularly updated as new queries were received and processed. This meant that all national coding centres had prompt access

to an additional source of advice about responses that had been found problematic in some sense. Coding supervisors in all countries found this to be a particularly useful resource though there was considerable variation in the number of queries that they submitted.

Field trial outcomes

Extensive analyses were conducted on the field trial cognitive item response data. These analyses have been reported elsewhere, but included the standard *ACER ConQuest*® item analysis (item fit, item discrimination, item difficulty, distractor analysis, mean ability and point-biserial correlations by coding category, item omission rates, and so on), as well as analyses of gender-by-item interactions and item-by-country interactions. On the basis of these critical measurement statistics, it was recommended that seven new items be removed from consideration for the main survey. In addition, the coding of partial credit items was reviewed. In some cases, the collapsing of categories was recommended. Consortium members also examined the items showing language DIF and considered whether issues in translating the item might be the source of the language DIF. Minor modifications were made to a small number of items (in either English or French source versions) if translation issues were thought to have contributed to an item showing language DIF. The parts of each complex multiple-choice item were also analysed separately and this led to some parts being dropped though the item itself was retained.

National review of field trial items

A further round of national item review was carried out, this time informed by the experience at national centres of how the items worked in the field trial in each country. A document, *Item Review Guidelines,* was produced to assist national experts to focus on the most important features of possible concern. In addition, NPMs were asked to assign a rating from 1 (low) to 5 (high) to each item to indicate its priority for inclusion in the main survey. About half of the countries completed this review of the field trial items. For digital reading, 14 of the 23 participating countries provided feedback on the field trial digital reading items – again via the online review system.

A comprehensive field trial review report also was prepared by all NPMs, for both the paper-based and computer-based assessments. These reports included a further opportunity to comment on particular strengths and weaknesses of individual items identified during the translation and verification process and during the coding of student responses.

MAIN STUDY

Main survey reading item selection

The Reading Expert Group (REG) met on 22-25 September 2008 in Melbourne to review all available material and recommend which reading items should be included in the main survey.

The REG members considered the pool of 205 print reading items (new items and items retrieved from the 2000 administration) that had been field trialled and had performed adequately in terms of psychometric quality, at initial review (seven items had previously been rejected by the Consortium as technically inadequate, on the basis of analysis of the field trial data). The 205 items were evaluated by the REG in terms of their substantive quality, fit to framework, range of difficulty, national centre feedback, and durability. Similarly, of the digital reading pool of 72 field trial items, 11 items were judged of insufficient technical quality to be considered for the main survey. The remaining 61 items were reviewed by the REG using a similar set of criteria to that used for the print item selection.

The selections in both cases had to satisfy the following conditions:

- the psychometric properties of all selected items had to be satisfactory;
- items that generated coding problems had to be avoided unless those problems could be properly addressed through modifications to the coding guides; and
- items given high priority ratings by national centres were to be preferred, and items with lower ratings were to be avoided.

In addition, the item set (in the case of print reading, the combined set of new and link items) had to satisfy these conditions as much as possible:

- the major framework categories had to be populated as specified in the reading literacy framework; and
- there had to be an appropriate distribution of item difficulties.

The REG made a preliminary selection of print reading units (including eight "easy" units), and then selected items from the agreed units. After the test developers had provided a summary of the preliminary selections, the REG made final adjustments to the recommended sets. The REG recommended that the print reading main survey pool be selected from a set comprising 28 trend items, 16 PISA 2000 link items, and 129 new items. The selection came from 20 sources (14 national centres and 5 Consortium groups) and was originally in 12 source languages. The selected material received an average rating from national centres on "priority for inclusion" of 3.81.

For digital reading, the REG recommended that the main survey items be selected from a set of 11 units comprising 46 items. As noted earlier, the majority of material in the digital reading pool was generated by the test development centres, but two nationally submitted units were recommended for the main survey. The selected material received an average rating from national centres on "priority for inclusion" of 4.01.

The main survey item pools were presented during a meeting of NPMs in Sydney, Australia in September/October 2008.

Subsequently, for print reading, one new unit was dropped from the item pool as a result of NPM concerns about the appropriateness of its context in some cultures, and another unit that had not been included in the REG selection was reinstated, when a large number of NPMs expressed their disappointment at its exclusion. One other new unit was included to adjust for framework balance, and one further new unit, one unit retrieved from PISA 2000, and 29 single items from field trialled units recommended by the REG were omitted because of space considerations.

The numbers of new items, items retrieved from PISA 2000 and link items in the final selection for the main survey is shown in Table 2.8.

Table 2.8 Print reading main survey cognitive items

New items	94
PISA 2000 retrieved items	11
Link items	28[1]
Total	131

[1] Two items in the link set were omitted from the analysis of the main survey items because of poor reliability.

For digital reading, the NPMs endorsed the REG's recommended selection pool at their September meeting. Subsequently, two full units and 11 individual items from selected units were omitted from the main survey item pool because of space limitations. In total 29 digital reading items were included in the main survey.

Distributions of the print reading items, with respect to the major framework variables, are summarised in Table 2.9, Table 2.10 and Table 2.11.

Table 2.9 Print reading main survey items (item format by aspect)

	Access and retrieve	Integrate and interpret	Reflect and evaluate	Total
Multiple choice	6	38	8	52 (40%)
Complex multiple choice	3	6	1	10 (8%)
Closed constructed response	9	4	0	13 (10%)
Short response	10	1	0	11 (8%)
Open constructed response	3	18	24	45 (34%)
Total	**31 (24%)**	**67 (51%)**	**33 (25%)**	**131 (100%)**

Table 2.10 Print reading main survey items (item format by text format)

	Continuous	Mixed	Multiple	Non-continuous	Total
Multiple choice	36	4	2	10	52 (40%)
Complex multiple choice	6	1	0	3	10 (8%)
Closed constructed response	4	0	2	7	13 (10%)
Short response	4	1	0	6	11 (8%)
Open constructed response	31	1	1	12	45 (34%)
Total	**81 (62%)**	**7 (5%)**	**5 (4%)**	**38 (29%)**	**131 (100%)**

Table 2.11 Print reading main survey items (text type by aspect)

	Access and retrieve	Integrate and interpret	Reflect and evaluate	Total
Argumentation	5	16	9	30 (23%)
Description	10	11	9	30 (23%)
Exposition	8	23	9	40 (31%)
Instruction	6	1	4	11 (8%)
Narration	2	16	2	20 (15%)
Total	**31 (24%)**	**67 (51%)**	**33 (25%)**	**131 (100%)**

It was considered important that, other than differing in difficulty, the standard and easy booklets represented a similar alignment with the major framework variables in terms of distribution of items across categories. Percentage distributions of the print reading items across the standard and easy booklets, with respect to the major framework variables, are summarised in Table 2.12 to Table 2.15. The Full pool column shows the percentages of items in each category across the nine clusters used in the main survey for reading. The Standard test column shows the percentage per category for the seven clusters used in the standard booklets (Clusters R1, R2, R3a, R4a, R5, R6 and R7) and the Easy test column shows the parallel percentages for the clusters used in the easy booklets (Clusters R1, R2, R3b, R4b, R5, R6 and R7). The Target column shows the percentages aimed for in the framework.

Table 2.12 shows the distribution by percentage of items in the three categories of the aspect variable.

Table 2.12 Print reading main survey items in standard and easy tests (aspect %)

	Full pool	Standard test	Easy test	Target
Access and retrieve	24	23	24	25
Integrate and interpret	51	51	53	50
Reflect and evaluate	25	26	23	25
Total	**100**	**100**	**100**	**100**

Table 2.13 shows the distribution by percentage of items in the four categories of the text format variable.

Table 2.13 Print reading main survey items in standard and easy tests (text format %)

	Full pool	Standard test	Easy test	Target
Continuous	62	61	63	60
Mixed	5	7	6	5
Multiple	4	5	1	5
Non-continuous	29	27	30	30
Total	**100**	**100**	**100**	**100**

For the text format variable, efforts to reach the targets were concentrated on continuous and non-continuous, on which it was anticipated that reporting subscales might be built.

Table 2.14 shows the distribution by percentage of items in the five categories of the *text type* variable that were used in the print reading pool (no items were categorised as *transaction* by text type).

Table 2.14 Print reading main survey items in standard and easy tests (text type %)

	Full pool	Standard test	Easy test	Target
Argumentation	23	19	20	(no target)
Description	23	19	25	(no target)
Exposition	31	36	32	(no target)
Instruction	8	11	7	(no target)
Narration	15	16	16	15
Total	**100**	**100**	**100**	

For the text type variable, some sampling across the categories was sought, with a target percentage set only for *narration*.

Table 2.15 shows the distribution by percentage of items in each of the four categories of the *situation* variable.

Table 2.15 Print reading main survey items in standard and easy tests (situation %)

	Full pool	Standard test	Easy test	Target
Educational	27	27	27	28
Occupational	17	18	19	16
Personal	27	31	24	28
Public	29	24	30	28
total	**100**	**100**	**100**	**100**

Distributions of the digital reading items, with respect to the major framework variables, are summarised in Table 2.16 to Table 2.18. Table 2.16 shows the distribution of items by *aspect* and *item format*.

Table 2.16 Digital reading main survey items (item format by aspect)

	Access and retrieve	Integrate and interpret	Reflect and evaluate	Complex	Total
Multiple choice	7	9	2	0	18 (62%)
Complex multiple choice	0	1	0	2	3 (10%)
Open constructed response	0	0	4	4	8 (28%)
Total	**7 (24%)**	**10 (34%)**	**6 (21%)**	**6 (21%)**	**29 (100%)**

Digital reading introduces a unique variable for text, *environment*, which has two main categories: *authored* and *message-based*. A few items are based on texts representing both types of environment. These are categorised as *mixed*.

Table 2.17 shows the distribution of items by *environment* and *text format*.

Table 2.17 **Digital reading main survey items (environment by text format)**

	Authored	Message-based	Mixed	Total
Continuous	1	1	0	2 (7%)
Mixed	2	0	0	2 (7%)
Multiple	13	7	2	22 (76%)
Non-continuous	3	0	0	3 (10%)
Total	**19 (66%)**	**8 (28%)**	**2 (7%)**	**29 (100%)**

Table 2.18 shows the distribution of items by aspect and text type.

Table 2.18 **Digital reading main survey items (text type by aspect)**

	Access and retrieve	Integrate and interpret	Reflect and evaluate	Complex	Total
Argumentation	2	2	1	1	6 (21%)
Description	4	2	3	0	9 (31%)
Exposition	1	5	2	1	9 (31%)
Mixed	0	0	0	1	1 (3%)
Transaction	0	1	0	3	4 (14%)
Total	**7 (24%)**	**10 (34%)**	**6 (21%)**	**6 (21%)**	**29 (100%)**

The framework calls for sampling across text types, but no percentage targets were set.

Main survey mathematics items

Three clusters comprising a total of 24 units (35 items) were selected from the PISA 2003 main survey item pool. These were three intact clusters of the four clusters that had been administered in the main survey in PISA 2006. (The number of clusters for mathematics was reduced from PISA 2006 to PISA 2009 because, of the six cluster "slots" available for minor domains in both cycles, the REG had decided that in 2006 reading should administer exactly the same two clusters as it had administered in 2003 – thus allowing mathematics to fill the remaining four slots. However, in 2009, science and mathematics shared the six available slots equally.) The three clusters were selected to best represent the balance across framework variables.

Distributions of the mathematics items, with respect to the major framework variables, are summarised in Table 2.19, Table 2.20 and Table 2.21.

Table 2.19 **Mathematics main survey items (item format by competency cluster)**

	Reproduction	Connections	Reflection	Total
Multiple choice	5	1	3	9 (26%)
Complex multiple choice	0	6	1	7 (20%)
Closed constructed response	1	1	1	3 (9%)
Short response	2	6	0	8 (23%)
Open constructed response	1	4	3	8 (23%)
Total	**9 (26%)**	**18 (51%)**	**8 (23%)**	**35 (100%)**

Table 2.20 **Mathematics main survey items (item format by content category)**

	Space and shape	Quantity	Change and relationships	Uncertainty	Total
Multiple choice	2	3	1	3	9 (26%)
Complex multiple choice	1	2	2	2	7 (20%)
Closed constructed response	1	2	0	0	3 (9%)
Short response	1	4	1	2	8 (23%)
Open constructed response	3	0	5	0	8 (23%)
Total	**8 (23%)**	**11 (31%)**	**9 (26%)**	**7 (20%)**	**35 (100%)**

Table 2.21 **Mathematics main survey items (content category by competency cluster)**

	Reproduction	Connections	Reflection	Total
Space and shape	2	5	1	8 (23%)
Quantity	4	5	2	11 (31%)
Change and relationships	2	4	3	9 (26%)
Uncertainty	1	4	2	7 (20%)
Total	**9 (26%)**	**18 (51%)**	**8 (23%)**	**35 (100%)**

Main survey science items

Three clusters comprising a total of 18 units (53 items) were selected from the PISA 2006 main survey item pool, when science had been the major domain. These were not intact clusters, but they were intact units: no items or items parts that had been administered in PISA 2006 were omitted from the units selected for 2009. However, attitude items, which had been administered alongside cognitive units in PISA 2006, were not included.

Across the three clusters, units were selected that matched as closely as possible the 2006 distribution of competency classifications, knowledge classifications, item formats, range and distribution of item difficulties, difficulty by gender, and layout and cluster position.

Distributions of the science items, with respect to the major framework variables, are summarised in Table 2.22, Table 2.23 and Table 2.24.

Table 2.22 **Science main study items (item format by competency)**

	Identifying scientific issues	Explaining scientific phenomena	Using scientific evidence	Total
Multiple choice	4	8	6	18 (34%)
Complex multiple choice	6	7	4	17 (32%)
Closed constructed response	0	1	0	1 (2%)
Open constructed response	3	6	8	17 (32%)
Total	**13 (25%)**	**22 (42%)**	**18 (34%)**	**53 (100%)**

Table 2.23 **Science main study items (item format by knowledge type)**

	Knowledge of science	Knowledge about science	Total
Multiple choice	9	9	18 (34%)
Complex multiple choice	9	8	17 (32%)
Closed constructed response	1	0	1 (2%)
Open constructed response	7	10	17 (32%)
Total	**26 (49%)**	**27 (51%)**	**53 (100%)**

Table 2.24 **Science main study items (knowledge category by competency)**

	Identifying scientific issues	Explaining scientific phenomena	Using scientific evidence	Total
Physical systems	0	6	0	6 (11%)
Living systems	0	9	0	9 (17%)
Earth & space systems	0	7	0	7 (13%)
Technology systems	0	0	4	4 (8%)
Scientific enquiry	13	0	1	14 (26%)
Scientific explanations	0	0	13	13 (25%)
Total	**13 (25%)**	**21 (42%)**	**18 (34%)**	**53 (100%)**

Released items

The REG identified nine print reading units not included in the main survey that would be suitable for release as sample PISA reading units. One other unit was added to this set as a result of the NPM recommendations. In addition, four units of digital reading material from the field trial that were not included in the main survey were released. All of these units were included as an annex in the publication *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science* (OECD, 2010a).

No mathematics or science material was released after the 2009 field trial.

Despatch of main survey instruments

After finalising the main survey item selection, final forms of all selected items were prepared. This involved minor revisions to items and coding guides based on detailed information from the field trial, and the addition of further sample student responses to the coding guides.

For print reading, French translations of all selected items were then updated. Clusters of items were formatted as described previously, and booklets for were formatted in accordance with the main survey rotation design shown previously in Table 2.2. English and French versions of all items, item clusters and test booklets for the paper-based assessment were made available to national centres in three despatches, on 14 August (link clusters), 28 November (new reading units) and 19 December 2008 (new clusters and all booklets).

For digital reading, the English source version of the authored units was released for countries to make any necessary translation and adaptation changes on 21 November 2008. This release included both digital versions of the units and paper-based coding guides. The items were then arranged in clusters and test forms according to the main survey design shown in Table 2.3. The English source versions of the clusters and test forms were released on 16 December 2008.

Main survey coder training

Consolidated coding guides were prepared, in both English and (for the paper-based assessments) French, containing all the items that required manual coding. These were despatched to national centres on 23 January 2009. In addition, the training materials prepared for field trial coder training were revised with the addition of student responses selected from the field trial coder query service.

International coder training sessions for reading, mathematics and science were conducted in Brussels, Belgium in February 2009. All but four countries had representatives at the training meetings. As for the field trial, it was apparent at the training meeting that a small number of clarifications were needed to make the coding guides and training materials as clear as possible. Revised coding guides and coder training material for both paper-based assessments and the digital reading assessment were prepared and despatched early in March 2009.

Main survey coder query service

The coder query service operated for the main survey across the three test domains. Any student responses that were found to be difficult to code by coders in national centres could be referred to the Consortium for advice. The Consortium was thereby able to provide consistent coding advice across countries. Reports of queries and the Consortium responses were made available to all national centres via the Consortium website, and were regularly updated as new queries were received.

Review of main survey item analyses

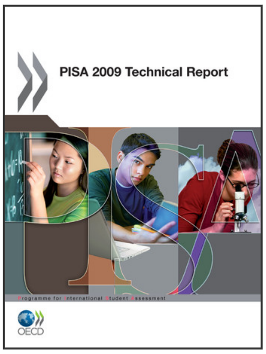
Upon reception of data from the main survey testing, extensive analysis of item responses was carried out to identify any items that were not capable of generating useful student achievement data. Such items were removed from the international data set, or in some cases from particular national datasets where an isolated problem occurred. Two reading items and one mathematics item were removed from the international data set.

Notes

1. This does not include the two items R219Q1E and R219Q1T that were deleted from the international analysis. These two items are not included in any of the succeeding discussion.

2. This does not include the mathematics item M305Q01 that was deleted from the international analysis.

3. Available at *www.pisa.oecd.org* > what PISA produces > PISA 2009 > PISA 2009 manuals and guidelines.



From:

PISA 2009 Technical Report

Access the complete publication at:

https://doi.org/10.1787/9789264167872-en

Please cite this chapter as:

OECD (2012), "Test Design and Test Development", in *PISA 2009 Technical Report*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264167872-3-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.