2

Test Design and Test Development

This chapter describes the test design for PISA 2012 and the processes by which the PISA Consortium, led by the Australian Council for Educational Research (ACER), developed the PISA 2012 paper-based tests for mathematics, reading and science, as well as for the international option, financial literacy. It also describes the design and development of the computer-based assessments of problem solving, mathematics and reading. In the following discussion, the term "mathematics" generally refers to the core paper-based mathematics assessment. The computer-based assessment of mathematics is referred to as "CBAM". The same applies in the case of reading: the computer-based assessment is referred to as the "digital reading assessment" (DRA). The PISA results reported publicly in December 2013 were from what is referred to as the PISA "Main Survey". This term is used to distinguish earlier developmental activities including those contributing to conduct of the "Field Trial" that occurred in 2011.

TEST SCOPE, DESIGN AND DEVELOPMENT

Test development for the PISA 2012 survey commenced in late 2009. Development proceeded through various processes and stages, slightly different for each of the cognitive domains in which test material was required, and culminating in the presentation to the PISA Governing Board (PGB) in October 2011 of a selection of items proposed for use in the 2012 Main Survey. This chapter presents the test design that governed the scope and structure of the PISA 2012 assessment, the development arrangements and approaches taken by ACER to produce the material required, and the processes of test development in each domain. Those domain-specific processes commenced with the specifications laid out in each assessment framework, and proceeded through the various stages of soliciting material for consideration, developing and refining that material to a finished form, seeking national feedback on the item developed, piloting and trialing draft material, and preparing materials fit and ready for use in the Main Survey.

The test design adopted for PISA 2012 specified the volume and arrangement of material needed in each domain that was to be tested (mathematics, reading, science, problem solving and financial literacy), and in each test mode that was to be employed (paper-based and computer-based). Those specifications required the development of sets of items (referred to as "item clusters") in each test domain, each of which would need to occupy a defined amount of test time. The specifications also determined how the item clusters would be arranged in test booklets (for the paper-based components) and in test forms (for the computer-based components).

Paper-based assessment design: mathematics, reading, science, financial literacy

The standard Main Survey items for mathematics, reading and science were to be compiled in thirteen item clusters (seven mathematics clusters, three reading 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, hence two hours of test time. Clusters labelled PM1, PM2, PM3, PM4, PM5, PM6A and PM7A denote the seven paper-based standard mathematics clusters, PR1 to PR3 denote the paper-based reading clusters, and PS1 to PS3 denote the paper-based science clusters.

PM1, PM2 and PM3 were the same three mathematics clusters as those administered in 2009, and the remaining clusters would comprise new material. Two of the three reading clusters were intact clusters used in 2009. The remaining reading cluster was based on a cluster used in 2009 but with one unit substituted. The substitution was made after the 2010 oil spill in the Gulf of Mexico rendered a unit about the idyllic nature of the Gulf unusable. The three science clusters were intact clusters used in PISA 2009.

The cluster rotation design for the standard booklets in the Main Survey corresponds to designs used in previous PISA surveys and is shown in Figure 2.1.

This is a balanced incomplete block design. Each cluster (and therefore each test item) appears in four of the four-cluster test booklets, once in each of the four possible positions within a booklet, and each pair of clusters appears in one (and only one) booklet. An additional feature of the PISA 2012 test design is that one booklet (booklet 12) is a complete link, being identical to a booklet administered in PISA 2009.

Each sampled student was randomly assigned to 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, typically of five minutes duration. Matters such as these on the administration of test sessions are described in more detail in Chapter 6.

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

■ Figure 2.1 ■

Cluster rotation design used to form standard test booklets for PISA 2012

Booklet ID	Cluster			
B1	PM5	PS3	PM6A	PS2
B2	PS3	PR3	PM7A	PR2
B3	PR3	PM6A	PS1	PM3
B4	PM6A	PM7A	PR1	PM4
B5	PM7A	PS1	PM1	PM5
B6	PM1	PM2	PR2	PM6A
B7	PM2	PS2	PM3	PM7A
B8	PS2	PR2	PM4	PS1
B9	PR2	PM3	PM5	PR1
B10	PM3	PM4	PS3	PM1
B11	PM4	PM5	PR3	PM2
B12	PS1	PR1	PM2	PS3
B13	PR1	PM1	PS2	PR3

as many items as the other booklets, with about 50% of the items being mathematics items (cluster PMUH), 25% reading (cluster PRUH) and 25% science (cluster PSUH). The items were selected from the Main Survey items taking into account their suitability for students with special educational needs, using criteria established in the lead-up to the PISA 2003 survey through consultation with the OECD Working Group on students with special educational needs.

In PISA 2012, as in PISA 2009, some countries were offered the option of administering an easier set of booklets whilst still providing an assessment that would generate results that are fully comparable to those from every other PISA participant, leading to an expanded booklet design incorporating material for both the standard PISA implementation and an implementation using the easier booklets. The offer was made to countries that had achieved a mean scale score in reading of 450 or less in PISA 2009, and to new countries that were expected – judging by their results on the PISA 2012 Field Trial conducted in 2011 – 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. 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 proficiency in mathematics. For countries that selected the easier set of booklets two of the standard mathematics clusters (PM6A and PM7A) were replaced with two easier mathematics clusters (PM6B and PM7B). Apart from level of difficulty, the sets of items in the standard and easier clusters were matched in terms of major framework characteristics to ensure that whichever set of items were taken in a particular country, the framework specifications were met. The other eleven clusters (five clusters of mathematics items, three clusters of reading items and three clusters of science items) were administered in all countries.

Although only two of the clusters differed for standard and easier administration, the cluster rotation in the booklets (where each cluster appears four times, once in each of the possible positions in the four-cluster booklets) means that more than half of the booklets are affected by the existence of these 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, as shown in Figure 2.2 in the full test design used in the paper-based component of the Main Survey including the optional components. The only difference between the two sets of booklets was that for some countries, booklets 1 to 7 (those containing clusters PM6A and PM7A) were replaced with booklets 21 to 27 (with the easier clusters PM6B and PM7B as substitutes).

In PISA 2012, an assessment of financial literacy was offered as an international option. Countries participating in this option administered an additional four booklets, each containing the two clusters of financial literacy items (denoted PF1 and PF2) as well as one cluster of mathematics material (cluster PM5) and one cluster of reading material (PR2). As with the core domains, a special one-hour booklet, referred to as the FLUH booklet (Financial Literacy Une Heure booklet), was prepared for use in schools catering for students with special needs. This booklet consisted of one cluster of financial literacy material (denoted PFUH), and one cluster of mathematics material (denoted PMUH). The items were selected from the Main Survey items taking into account their suitability for students with special educational needs. Countries administering the financial literacy assessment implemented booklets 71-74 (in addition to booklets 1-13 if administering the standard booklets, or 8-13 and 21-27 if administering the easier set of booklets).

■ Figure 2.2 ■

Cluster rotation design used to form all test booklets for PISA 2012

Booklet ID	Cluster				Standard booklet set	Easier booklet set
B1	PM5	PS3	PM6A	PS2	Y	
B2	PS3	PR3	PM7A	PR2	Y	
B3	PR3	PM6A	PS1	PM3	Y	
B4	PM6A	PM7A	PR1	PM4	Y	
B5	PM7A	PS1	PM1	PM5	Y	
B6	PM1	PM2	PR2	PM6A	Y	
B7	PM2	PS2	PM3	PM7A	Y	
B8	PS2	PR2	PM4	PS1	Y	Y
B9	PR2	PM3	PM5	PR1	Y	Y
B10	PM3	PM4	PS3	PM1	Y	Y
B11	PM4	PM5	PR3	PM2	Y	Y
B12	PS1	PR1	PM2	PS3	Y	Y
B13	PR1	PM1	PS2	PR3	Y	Y
B20 (UH)	PMUH	PRUH/PSUH				
B21	PM5	PS3	PM6B	PS2		Y
B22	PS3	PR3	PM7B	PR2		Y
B23	PR3	PM6B	PS1	PM3		Y
B24	PM6B	PM7B	PR1	PM4		Y
B25	PM7B	PS1	PM1	PM5		Y
B26	PM1	PM2	PR2	PM6B		Y
B27	PM2	PS2	PM3	PM7B		Y
B70 (FLUH)	PFUH	PMUH				
B71	PF1	PF2	PM5	PR2		
B72	PF2	PF1	PR2	PM5		
B73	PM5	PR2	PF1	PF2		
B74	PR2	PM5	PF2	PF1		

As was mentioned earlier, material used to populate the design-included item clusters that originated in earlier PISA surveys, included again here to facilitate the linking of ability estimates across survey administrations, as well as new mathematics material needed to support the expansion of mathematics to "major domain" status for the PISA 2012 administration.

Computer-based assessment design: problem solving, mathematics, reading

For PISA 2012, a computer-based assessment of problem solving was included as part of the core assessment, which was taken up by about two-thirds of participating countries. Whilst the PISA Governing Board had wished to introduce the problem solving component for the PISA 2012 survey, a number of countries for a variety of technical and other reasons were not able to meet this wish. Nevertheless, problem solving continued to be referred to as a core component of the assessment.

In addition, countries were offered assessments of computer-based mathematics (CBAM) and reading in a digital environment (DRA). The latter two were offered together, in an assessment of computer-based literacies (CBAL). Countries implementing any part of the assessment on computer would either administer the assessment of problem solving only, or, assessments of all three of problem solving, CBAM, and DRA. They could not choose to administer CBAL while opting out of the assessment of problem solving.

The Main Survey items for the problem solving assessment were to populate four item clusters with each cluster representing 20 minutes of test time. For the countries administering the problem solving assessment as their only computer-based component, the test design specified that items would be presented to students in eight test forms, with each form being composed of two clusters according to the rotation design shown in Figure 2.3. The labels CP1 to CP4 denote the four computer-based problem solving clusters.

Each sampled student was randomly assigned one of the eight forms, which meant each student undertook 40 minutes of testing.

■ Figure 2.3 ■

Main Survey test design for countries participating in problem solving only

Form	Cluster 1	Cluster 2
31	CP1	CP2
32	CP2	CP3
33	CP3	CP4
34	CP4	CP1
35	CP2	CP1
36	CP3	CP2
37	CP4	CP3
38	CP1	CP4

Main Survey items for the CBAM and DRA were to populate four and two item clusters respectively, with each cluster representing 20 minutes of test time. For the countries administering the problem solving assessment together with the CBAL, the design specified that items would be presented to students in 24 test forms, with each form being composed of two clusters according to the rotation design shown in Figure 2.4. The labels CM1 to CM4 denote the four computer-based mathematics clusters, and CR1 and CR2 denote the two digital reading clusters.

Each sampled student was randomly assigned one of the 24 forms, which meant each student undertook 40 minutes of testing.

■ Figure 2.4 ■

Main Survey test design for countries participating in problem solving and CBAL

Form	Cluster 1	Cluster 2
41	CP1	CP2
42	CR1	CR2
43	CM3	CM4
44	CP3	CR1
45	CR2	CM2
46	CM1	CP4
47	CR2	CR1
48	CM2	CM1
49	CP3	CP4
50	CM4	CR2
51	CP1	CM3
52	CR1	CP2
53	CM1	CM3
54	CP4	CP1
55	CR1	CR2
56	CP2	CM4
57	CR2	CP3
58	CM2	CR1
59	CP2	CP3
60	CM4	CM2
61	CR2	CR1
62	CM3	CP1
63	CR1	CM1
64	CP4	CR2

Domain definitions, and item design: The 2012 assessment frameworks

The material needed to fulfil the design requirements had to satisfy the domain definitions and specifications within the relevant assessment framework. For each PISA subject domain, an assessment framework is produced to guide the PISA assessments in accordance with the policy requirements of the PISA Governing Board. 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 mathematical literacy had been partially developed

for the first PISA administration in 2000, and more fully articulated for PISA 2003 when mathematics was the major test domain for the first time, the PISA 2012 work began with a review of the existing framework at the initial meeting of the Mathematics Expert Group (MEG) in October 2009. That review and subsequent development work was carried out jointly by ACER and Achieve, the organisations appointed by the PISA Governing Board to jointly revise the mathematics framework for PISA 2012, in accordance with a development plan and timeline adopted by the PGB at its November 2009 meeting. Work on mathematics framework development commenced in October 2009 and continued through to adoption of the framework by the PISA Governing Board in November 2010.

A preparatory step in this development process was a survey of mathematical content standards applying in a range of relatively high-performing OECD countries, carried out by Achieve. Countries in that analysis included Australia, Belgium, Canada (Alberta), Finland, Ireland, Japan, Korea, New Zealand, and the United Kingdom. Achieve also analysed the previous frameworks and co-ordinated an extensive consultation process on the revised framework with experts from a range of countries as the PISA 2012 framework was under development. That consultation included consideration of responses to a detailed survey instrument, with responses from over 80 individuals (largely mathematicians and mathematics educators) from 34 countries participating in PISA. Several changes were proposed to the framework: *(i)* a revised definition of mathematical literacy was proposed and successively refined; *(ii)* the ways in which mathematical content was conceptualised and described underwent considerable revision over several drafts; *(iii)* the definition and description of mathematical processes were very substantially changed, resulting in a configuration of processes that would underpin a new set of reporting dimensions for PISA mathematics outcomes; and *(iv)* the contexts within which opportunities for students to express their levels of mathematical literacy would be provided were also reviewed and revised. Extension of the framework to incorporate a computer-based assessment option was developed, and a set of background variables that would be of particular interest was identified for mention in the framework. Revised framework drafts were presented in 2010 to successive meetings of the PGB which adopted a final version in 2011. An external validation of the item pool was implemented by Achieve to support the PGB's consideration of the items proposed to be used in the PISA 2012 survey instruments. Achieve engaged a team of mathematics experts to carefully review the items, and provided an independent external judgement about the fit of each item to the new framework.

The reading and science frameworks were unchanged in PISA 2012. However, new frameworks for two components of the PISA 2012 survey, the computer-based assessment of problem solving, and the assessment of financial literacy, were developed by ACER and its collaborators so that they could be adopted and published as part of the consolidated framework publication for PISA 2012.

Computer delivery was fundamental to the conception of problem solving in PISA 2012. It enabled *interactive* problems – problems in which exploration is required to uncover undisclosed information (Ramalingam et al., 2014) – to be included in a large-scale international assessment for the first time. In developing these problems, the emphasis was on everyday problem situations that often arise when interacting with an unfamiliar device (such as a ticket vending machine, air-conditioning system, or mobile phone) for the first time. Some of these devices, such as vending machines, were modelled as finite state machines (Buchner and Funke, 1993; Funke, 2001), that is, systems with a finite number of states, input signals and output signals. The system's next state is determined by its current state and the specific input signal selected by the user. Other problem situations, such as controlling an air conditioner, involved manipulating input and output variables that are related in some causal way. These situations were implemented as "MicroDYN" units (Greiff et al., 2013; Wüstenberg et al., 2012).

A particular challenge in task development for the PISA 2012 problem-solving assessment arose from the requirement to construct problems that did not need any particular domain-based knowledge for their solution, and with which students were not already familiar. This was intended to ensure that the focus was on measuring the cognitive processes involved in problem solving in a way more or less uncontaminated by the specific domain-based knowledge students had previously acquired through their other studies. This approach constitutes a major difference from the assessment of the other core domains in PISA (reading, mathematics and science), in which the assessments are constructed so that expert knowledge in the domain is required, indeed forms a main target of the assessment. For the assessment of problem solving, wherever possible low-verbal and non-verbal information was used in describing problems, hence minimising potential dependence on reading literacy skills, and only a basic level of mathematical and scientific knowledge was involved. In reality, ensuring that problems are equally unfamiliar to students is impossible at the individual level, but could perhaps be achieved across countries by presenting a variety of contexts so that no one group was consistently advantaged or disadvantaged in this way.

The assessment of financial literacy was an international option for PISA 2012. Development of the framework was overseen by a group that represented the expertise of the OECD Financial and Enterprise Affairs Directorate and the international experts who had been advising the Directorate in its efforts to promote financial education around the world.

The assessment framework drew heavily on work of the OECD-sponsored International Network on Financial Education (INFE), established in 2008, as well as on that of individual researchers at the national level. Like other PISA literacy domains, the financial literacy assessment framework set out ways of measuring the proficiency of 15-year-olds in demonstrating and applying knowledge and skills, while recognising that certain limitations had to be taken seriously given the enormous variation among OECD countries in the legislative, regulatory and practical approaches taken to financial matters. Key concepts to be included were the **content** of financial literacy (identified as *money and transactions*, *planning and managing finances*, *risk and reward* and *financial landscape*) and essential **processes** (*identify financial information*, *analyse information in a financial context*, *evaluate financial issues* and *apply financial knowledge and understanding*). The framework also identified four **contexts** in which the financial literacy of 15-year-olds should be demonstrated: *education and work*, *home and family*, *individual* and *societal*.

The items for the 2012 financial literacy assessment were developed by ACER and presented to the financial literacy expert group for feedback. It was the role of the expert group to ensure that the items developed matched the financial literacy framework that was being developed in parallel at the time. Advice was also sought from the expert group on whether the items were suitably aligned with the varied financial systems of the different countries taking part in the assessment. Due to time constraints, the various National Centres were unable to provide robust feedback on the items, but they were able to alert ACER to items that were inconsistent with their own financial systems and practices. Many of the items formed part of small units (consisting of between two and four items) whereas other items were stand-alone questions. In total, 81 items were included in the Field Trial and 40 items were included in the Main Survey. The items comprised simple vocabulary and no more than basic mathematics so as not to disadvantage those students with low reading and mathematics abilities.

One of the greatest challenges of item development for the financial literacy assessment was creating scenarios that applied equally to students from the different participating countries. For example, the financial consumer's relationship with credit cards and credit services varies widely between countries, and so the scenarios developed around credit had to be non-specific, ensuring that different countries' students were neither advantaged nor disadvantaged by the items. Similarly, items involving taxation had to be fairly generic to reflect the different taxation systems used in the different countries. Items involving value-based judgements were generally avoided as it was not considered sensible to use items to assess a student's attitude to saving and spending, noting that what may be a "sound" financial decision for the majority of people may not be the case for certain individuals in certain circumstances.

Another problem that had to be resolved with the expert group was the degree of financial knowledge and skills expected of a 15-year-old to "enable participation in economic life" (part of the framework's definition of financial literacy). Many financial concepts are beyond the first-hand experiences of the typical 15-year-old, with scenarios like pension contributions far off the student's radar. Financial scenarios such as shopping and saving up for a large purchase are commonplace activities throughout all countries but relying on such basic scenarios would limit the efficacy of the assessment. Some participating countries already had in place financial education courses for students but many others did not, and the lack of consistency among those existing financial education frameworks meant that much of the assessment framework was developed with fewer models to draw on than domains such as mathematics and science.

In 2012, the framework was prepared for publication along with an extensive set of example items. All five PISA 2012 cognitive frameworks were published in *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy* (OECD, 2013).

Test development centres

Experience gained in the four 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 mathematics and problem solving items for PISA 2012, ACER drew on the resources of nine test development centres in culturally-diverse and well-known institutions, namely ACER, the University of Melbourne (both in Australia), aSPe (University of Liege, Belgium), DIPF (Deutschen Institut für Internationale Pädagogische Forschung), IPN (Leibniz-Insitute for Science and Mathematics Education) and

Heidelberg University (all three in Germany), NIER (the National Institute for Educational Policy Research, Japan), CRP-HT (the Centre de Recherche Public – Henri Tudor, Luxembourg), ILS (the Department of Teacher Education and School Research, University of Oslo, Norway) and ETS (Education Testing Service, United States). For financial literacy, all new test development was undertaken at ACER.

ACER co-ordinated the distribution of material for development across these centres, and managed the co-operative development processes in which the item writers in each centre engaged. The test development teams were encouraged to conduct initial development of items, including cognitive laboratory activities, in their local language. Translation to the OECD official languages (English and French) took place after items had reached a well-formed state.

Scope, volume and constraints

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 from one to five items assessing students' competencies and knowledge. 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.

For the paper-based assessment, 56 mathematics units comprising a total of 110[1] cognitive items were needed to provide approximately 270 minutes of testing time for mathematics in PISA 2012. The reading assessment consisted of 44 items (13 units), a subset of the 131 items used in 2009, 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 the same as those used in 2009. The optional assessment of financial literacy consisted of 29 units, comprising a total of 40 items, representing 60 minutes of testing time (see Annex A).

The 110 cognitive mathematics items used in the Main Survey included 36 items from the 2003 test that had also been used for linking in 2006 and 2009. The remaining 74 items were newly developed for PISA 2012. The 74 new items were selected from a pool of 172 newly-developed items that were tested in a Field Trial conducted in all countries in 2011, one year prior to the Main Survey. The 40 items comprising the financial literacy assessment were newly-developed for PISA 2012, and were selected from a pool of 75 items that were similarly tested in a Field Trial conducted in 2011 in countries participating in this international option. There was no new item development for reading or science, as the design requirements could be met with existing secure material.

The problem solving assessment comprised sixteen units, with a total of 42 items, representing 80 minutes of testing time in total. These items were selected from a pool of 79 newly-developed problem solving items that were tested in a Field Trial conducted in all participating countries in 2011, one year prior to the Main Survey. The instrument for the CBAM comprised 15 units, with a total of 41 items, representing 80 minutes of testing time in total. These items were selected from a pool of 86 newly-developed computer-based mathematics items that were tested in a Field Trial conducted in all participating countries in 2011, one year prior to the Main Survey. As well as the item format types referred to in relation to the paper-based assessment items, additional variants of the selected response format type were used with items that involved, for example, selection from a drop-down menu, use of "drag and drop" and use of "hot spots".

The instrument for the DRA assessment consisted of 19 items, based on 6 units, representing 40 minutes of testing time. The digital reading items were selected from the 29 items used in the DRA in PISA 2009.

In each of the computer-based assessments, 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 continue with the item they had been viewing.

The assessment items for problem solving and computer-based mathematics each make use of only one screen of stimulus material, 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. In the case of the DRA, then, 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.

Item formats employed with paper-based cognitive items were either selected response or constructed response. Selected response 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 two or more possible responses (yes/no, true/false, correct/incorrect, etc.). Constructed response items were of two broad types. Constructed response manual items required limited manual input by trained coders at the stage of processing student responses. They required students to construct a numeric response within very limited constraints, or only required a word or short phrase as the answer, and coders later to assign each response to the predefined response categories. Constructed response expert items required the use of trained expert coders to interpret observed student responses and assign them to one of the defined response categories. These items required a response to be generated by the student, with a range of possible full-credit answers.

For the computer-based cognitive items, two additional item formats were employed. The first, constructed response auto-coded, included any item in which students constructed a non-text based response. This might be done, by, for example, highlighting segments of map to show an optimal route, or dragging and dropping an object from one point to another. As the name suggests, scoring rules were defined for such items so that they could be coded automatically. The other new response format was "selected response variations". These included any item in which the student selected a response that was not multiple-choice or complex multiple-choice. This item type included drop down menu items where either a) there was more than one drop down menu; b) there was more than one possible correct response; or c) where more than one choice could be made. For example, select the best two responses from the following list.

Pencils, erasers, rulers, and in some cases calculators, would be provided to students undertaking the PISA assessment. 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 students accustomed to their use.

Development timeline and processes

Planning for mathematics item development began in September 2009, with preparation of material for a two-day meeting of test developers from each test development centre, which was held in Offenbach on 19-21 October, 2009. The meeting had the following purposes:

- to become familiar with the issues under consideration by ACER and Achieve in revising the mathematics framework for PISA 2012, especially the implications of possible changes 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 four 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.

The meeting reviewed documentation prepared by ACER to guide all parts of the process for the development of cognitive items: the calling for submissions from participating countries, writing and reviewing items, carrying out cognitive laboratory activities and pilot tests of items and conducting an extensive Field Trial, producing final source versions of all items in both English and French, preparing coding guides and coder training material, and selecting and preparing items for the Main Survey, all in time to distribute material to PISA National Centres in each participating country well in advance of the commencement of the Main Survey in March 2012. The main phase of test development finished when the items were distributed for the Field Trial in December 2010. 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 2011, after which items were selected for the Main Survey and distributed to countries in December 2011.

The material from which the new mathematics items were developed originated from three main sources. First, the National Centres from participating countries submitted a large number of items or ideas for items, some 500 in total including about 400 intended for paper-based delivery and about 50 intended for computer delivery. Material was

submitted by twenty different National Centres (Canada, Colombia, the Czech Republic, France, Greece, Israel, Italy, Korea, Macao-China, Mexico, the Netherlands, Portugal, Serbia, Shanghai-China, Singapore, Spain, Switzerland, Chinese Taipei, Uruguay and the United States). Second, the members of the Mathematics Expert Group and Consortium staff working with that group contributed a small pool of items, many of which were designed to expand the volume of relatively easy material available for selection. Third, the teams of professional item writers engaged by ACER to develop material provided a significant volume of original material, in addition to the development work those teams carried out to refine submitted material.

The development timeline for the problem solving items was similar to that for mathematics, although heavy involvement of test development centres outside ACER occurred at a slightly later point in the development process. The items for the PISA 2012 problem-solving assessment came from two sources: the PISA international Consortium and national submissions. After initial development work by the test development centres, the Problem-Solving Expert Group that developed the PISA 2012 framework reviewed materials to ensure that they reflected the defined construct of problem-solving competence. Small-scale cognitive laboratory activities were conducted, and the items were reviewed by National Centres and field tested.

First phase of development

Typically, the following steps were taken in the first phase of the development of mathematics items. A similar process, simplified and shortened in some cases, was followed in the other (minor) domains for which new item development was needed. 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

At the early stages of test development, test developers in each of the Consortium test development centres found potential material and exchanged it with one or more other centres (in English translation if necessary) to ascertain whether colleagues agreed that it was worth developing further, or they worked with material that had originated in national item submissions that had been assigned to them for development. The material was formatted even at this early stage in a manner similar to that planned for the final presentation.

For material that was 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 paper-based 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 computer-based items, all cognitive interviews were conducted individually, using either audio-recording of responses or dual administration, with one researcher interacting with the student and a second 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 for several of the items, 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 individual items or even whole 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.

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 items. The document *Item Development for PISA 2012 and Item Submission Guidelines* was distributed to PISA 2012 National Project Managers in March 2010.

The guidelines described the scope of the item development task for PISA 2012, the arrangements for national submissions of items and provided sample items. In addition, the guidelines contained a detailed discussion of item requirements and an overview of the full item development process for PISA 2012.

To assist countries in submitting high quality and appropriate material, ACER conducted a one-day mathematics item development workshop for interested National Centres at the end of the first meeting of National Project Managers (NPMs) for PISA 2012, in March 2007. It was attended by individuals from most National Centres. The due date for national submission of items was 31 May 2010 for problem solving, and 1 June 2010 for mathematics, as late as possible given Field Trial preparation deadlines. Items could theoretically be submitted in any language, but in many cases the preliminary development work that occurred in country concluded with the preparation of an English language version prior to submission. 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. For mathematics, 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.

Approximately 450 items were submitted by PISA National Centres for consideration by the international contractor's test development teams. These items came from about 20 different countries. Some submitted units had already undergone significant development work. Others were in a less developed state. All submitted material was initially reviewed by the

test development co-ordinator at ACER, to check for consistency with the framework, to identify material that was repetitive (for example, to identify instances where two different National Centres had submitted material that was very similar, or that was too similar to material already in development) or that may have been unsuitable for other reasons (such as being too ephemeral, or sensitive on cultural grounds). Where material was deemed suitable at this initial screening stage, it was assigned to one of the test development teams, after which the processes described earlier were applied.

National review of items

In July 2010, National Project Managers (NPMs) were given a set of item review guidelines to assist them in reviewing cognitive items and providing feedback, using an online review and feedback system that was developed by ACER for this purpose. Bundles of items were made available progressively through 2010 as item development proceeded, with item bundles being released in March, April, July, and two in August 2010. 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 mathematics. Corresponding feedback categories were used for the other domains. 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. For items intended for computer delivery (CBAM and problem solving), feedback was also sought on the likely demands related specifically to general computer use and familiarity that would be essentially unrelated to the cognitive objectives of the items.

For each bundle, a series of reports was generated summarising the feedback from National Project Managers. The feedback frequently resulted in useful input to the international contractor's test development teams in its task of further revising 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 PISA MEG that took place in 2010 and 2011.

In addition, as mentioned earlier Achieve conducted an independent external validation study in relation to the mathematics items selected for use in the Field Trial, to assess the extent to which they were a proper reflection of the objectives and constraints specified in the mathematics framework. The conclusion in the report of the validation study was:

> *"... that the items represent the framework well, and cover the mathematics expected of 15-year-olds at an appropriate breadth and depth. Also, assuming the selection of operational items from this field test pool addresses concerns voiced by the external validation panel, they agreed that PISA 2012 will assess the construct of mathematical literacy as defined in the framework."*

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, ensuring synchronisation of the two source 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.

Field testing

The PISA Field Trial was carried out in all countries with the implementation occurring for the majority of countries in the first half of 2011. 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 2011, the test development centres, the mathematics, problem solving and financial literacy expert groups and National Centres reviewed the Field Trial data to support the identification of a proposed selection of Field Trial items for the Main Survey.

Field Trial item selection

A total of a 474 mathematics items (344 paper-based and 130 computer-based) were circulated to National Centres for review from early March to late August 2010. Seventy-four of those (65 paper-based and 9 computer-based) had originated from national submissions.

From that pool of 474 items, 172 paper-based items were selected to supplement the pre-existing 36 link items, and 86 computer-based items were selected, for inclusion in the Field Trial. The selection of those items took into account a number of factors: the rating of items by national experts (their priority for inclusion) as part of their review of the item bundles, other item feedback from National Centres bearing on item quality and acceptability, the preferences of expert group members based largely on the fit of items to the objectives and definitions of the framework, data derived from cognitive laboratories and small-scale pilot activities including data on the expected difficulty of items, and the need to balance the selection against the framework's test specification.

A similar selection process occurred for the problem solving items, where 79 items were selected for inclusion in the Field Trial; and likewise for financial literacy, where 75 items were selected.

For the paper-based reading and science components of the Field Trial, material was used in intact clusters from previous PISA administrations (with the exception of one reading unit replaced as mentioned earlier); and likewise for the digital reading component, which used intact material from the 2009 digital reading assessment.

Field Trial design

Paper-based assessment

The Field Trial design for the paper-based assessment comprised 17 clusters of mathematics items (denoted PM1 to PM17), 3 clusters of reading items (PR1 to PR3) and 3 clusters of science items (PS1 to PS3).

Clusters PM1, PM2 and PM3 were intact clusters that had been used in PISA 2003, 2006, and 2009 comprising 36 link items (in 25 units). The 172 new mathematics items (from 62 units) were allocated to 14 clusters, PM4 to PM17.

PR1 and PR2 were two intact reading clusters from PISA 2006 and PR3 was an almost intact cluster from 2006 but with one three-item unit inserted in place of material that had to be replaced. These three clusters comprised 44 items (13 units). PS1, PS2 and PS3 were 3 intact science clusters comprising 53 items (18 units) selected from the 2006 survey.

Material for the optional financial literacy component comprised 75 items placed in four clusters. In addition, the Field Trial design included a one-hour test booklet comprising one mathematics cluster, a half cluster of reading material and a half cluster of science material, for special educational needs students. Items in these clusters were selected taking into account their suitability for students with special educational needs.

Ten 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 mathematics clusters appeared once in the first half of a booklet and once in the second half, in booklets 1 to 8, and were administered in all participating countries. The mathematics, reading and science link material appeared in booklets 9 and 10; these booklets were administered only in countries participating in PISA for the first time in 2012. Figure 2.5 shows the Field Trial design for the paper-based assessment.

Two one-hour booklets were administered in the Field Trial to support the testing of sampling and operational procedures in schools having students with special educational needs, one for students in the regular sample (labelled BUH in Figure 2.5), and one for students in the sample for the financial literacy international option (labelled BFUH).

The booklets used were identical to those that had been used in the PISA 2009 test booklet rotation design. Booklet BUH comprised a reading cluster labelled in Figure 2.5 as PRUH, and two half clusters (one for each of mathematics and science) labelled as PMUH and PSUH. Exactly the same clusters were used in the BFUH booklet.

■ Figure 2.5 ■

Allocation of item clusters to test booklets for Field Trial

Booklet ID	Cluster				Booklet set for:
B1	PM4	PM12	PM13	PM6	All participating countries
B2	PM5	PM13	PM14	PM7	All participating countries
B3	PM6	PM14	PM15	PM8	All participating countries
B4	PM7	PM15	PM16	PM9	All participating countries
B5	PM8	PM16	PM17	PM10	All participating countries
B6	PM9	PM17	PM1	PM11	All participating countries
B7	PM10	PM1	PM2	PM4	All participating countries
B8	PM11	PM2	PM12	PM5	All participating countries
B9	PM3	PS1	PS2	PS3	Only new countries
B10	PR1	PR2	PR3	PM3	Only new countries
BFL1	PF1	PF2	PF3	PF4	
BFL2	PF4	PF3	PF2	PF1	
BUH	PRUH	PMUH/PSUH			
BFUH	PRUH	PMUH/PSUH			

Computer-based assessment

The 86 computer-based mathematics items were arranged in eight clusters each designed to occupy 20 minutes of test time, and these were administered in pairs in eight test forms, hence each form occupied 40 test minutes.

The 79 Field Trial items for problem solving were also arranged in eight twenty-minute clusters, and these were also administered in pairs in eight test forms.

Two twenty-minute clusters of computer-based reading material were formed from the 18 items, and delivered in two test forms.

Dispatch of Field Trial instruments

Field Trial instruments were dispatched to PISA National Centres in stages during the period from late October to December 2010 as they reached their final form.

Final versions of material for computer delivery were released in the online translation management system in October 2010. Final English and French paper-based source versions of the new mathematics Field Trial units were distributed to National Centres in two batches, the first in November 2010 (along with the financial literacy material), and the second in early December 2010. All consolidated final source versions of booklets (in English and French) and forms (in English) were distributed on 22 December 2010. All material could also be downloaded from the PISA website from the time of dispatch.

As material became available, National Centres 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 dispatch of items to be included in the Field Trial, various documents and materials were prepared to assist in the training of personnel who would lead the coding of student responses in each PISA country. International coder training sessions for mathematics, reading, science, problem solving and financial literacy were conducted in February 2011. 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 defined different kinds of possible responses to each item, 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. Corresponding training

material was also prepared for the computer-based components. Coding of response to computer-based items was carried out in an online coding system developed for the purpose. Explanatory material guided the use of the system as well as showing how manually coded items should be treated, for each of problem solving, mathematics and reading.

Countries sent representatives to the training sessions. 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 dispatch of coding guides and training materials in March 2011. 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 to fill the role of expert coder.

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 as to the code most appropriate to a particular observed item response, National Centres were able to submit queries by e-mail to the query service, and these 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 periodically 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. Over successive PISA administrations, the accumulated coder queries have provided an excellent source of additional examples for the coding guides and training materials.

Field Trial outcomes

Extensive analyses were conducted on the Field Trial cognitive item response data, and 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. In reviewing those statistics, for example, response categories needed to be well ordered according to the average abilities of students giving each response; the point-biserial correlation for the key category should be positive, and for the other categories much smaller or negative; the fit of items should be near to 1. These data would be vital information to be used in the selection of items for use later in the Main Survey. In addition, the coding of partial credit items was reviewed. In some cases, the collapsing of categories was recommended.

Consortium analysts routinely examined all items for evidence of Differential Item Functioning (DIF), whereby different subsets of the assessed population (for example, different gender groups, country or language groups) when matched for ability, found the items differentially difficult. Any such cases were carefully examined to determine whether wording, translation or other factors in the presentation of the item may have contributed, and if so whether the issue could be resolved through some minor adjustment of the item, or could not easily be resolved in which case the item was set aside as unsuitable for selection in the Main Survey item pool.

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 in the online item review system, this time informed by the experience at National Centres of how the items had worked in the Field Trial in each country. A document, *Item Review Guidelines*[2] 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. A high proportion of participating countries completed this review of the Field Trial items.

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 SURVEY PREPARATION, IMPLEMENTATION AND FOLLOW-UP

Main Survey item selection

The expert groups for mathematics, problem solving and financial literacy met in Melbourne in September 2011 to review all available material and recommend which items should be included in the Main Survey instruments.

The expert groups considered the pool of items (new items, and in the case of mathematics, items to be used to link 2012 outcomes to those of previous PISA administrations) that had been tested in the recent Field Trial and had performed adequately from a technical measurement perspective on the basis of the item statistics referred to in the previous section, and using criteria established in previous PISA survey analyses that are also referred to in Chapters 9 and 12 of this volume. The available items were evaluated by the expert groups in terms of their substantive quality, fit to framework, range of difficulty, National Centre feedback, and durability.

The selection of items to be proposed for inclusion in the Main Survey instruments had to satisfy the following conditions:

- the psychometric properties of all selected items had to be satisfactory (according to the criteria referred to above and in Chapters 9 and 12);
- items that generated coding problems in the Field Trial had to be avoided unless those problems could be properly addressed through modifications to the coding guides;
- items given high priority ratings by National Centres were to be preferred, and items with lower ratings were to be avoided;
- the major framework categories had to be populated as specified in the relevant framework; and
- there had to be an appropriate distribution of item difficulties, broad enough to generate useful measurement data at both extremes of the anticipated ability distribution of sampled students across all participating countries.

Recommended selections of items for mathematics (both the paper-based and computer-based components), problem solving and financial literacy were presented to a meeting of National Project Managers in October 2011 for their review and endorsement. Final recommendations were presented to the PISA Governing Board at its meeting in Israel in October 2011 for endorsement.

Characteristics of the mathematics item set used in the Field Trial, and the set used in the Main Survey, for both the paper-based and computer-based components, are summarised in Figure 2.6 showing the distribution of items in relation to the various categories specified in the framework.

■ Figure 2.6 ■

Mathematics item counts (Field Trial and Main Survey) by framework category

Framework category		Link items	New items			
			Paper-based		Computer-based	
			Field Trial	Main Survey	Field Trial	Main Survey
Content	Change and relationships	9	46	20	22	11
	Quantity	11	44	18	26	9
	Space and shape	9	42	18	19	12
	Uncertainty and data	7	40	18	19	9
Process	Formulate	11	42	22	16	9
	Employ	14	76	35	41	22
	Interpret	11	54	17	29	10
Context	Occupational	3	40	21	23	9
	Personal	5	50	16	22	13
	Public	14	42	15	17	11
	Scientific	14	40	22	24	8
Format type	Simple multiple choice	10	47	22	19	8
	Complex multiple choice	7	19	6	12	4
	Constructed response (automatic)				42	22
	Constructed response (expert)	8	60	23	9	4
	Constructed response (manual)	11	46	23		
	Constructed response (variations)				4	8

The item counts for mathematics, problem solving, reading, science and financial literacy (in each of the Field Trial and Main Survey) are presented in Figure 2.7.

■ Figure 2.7 ■

Item counts (Field Trial and Main Survey) by domain and delivery mode

Domain	Field Trial	Main Survey
Mathematics (paper-based)	208	110
Mathematics (computer-based)	86	41
Problem solving (computer-based)	79	42
Reading (paper-based)	44	44
Reading (computer-based)	18	18
Science (paper-based)	53	53
Financial literacy (paper-based)	75	40

Dispatch 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.

French translations of all selected items were then updated. For the paper-based material, clusters of items were formatted, and booklets were formed in accordance with the Main Survey rotation design shown previously in Figure 2.2. For the computer-based material, the release of units included both digital versions of the units, and paper-based coding guides. English and French versions of all material were made available to National Centres in several dispatches, on 2 September (link clusters), 24 November and 5 December (new paper and computer-based units) and 20 December 2011 (new clusters and all booklets).

Main Survey coder training

Consolidated coding guides were prepared, in both English and French, containing all the items that required manual coding. These were dispatched to National Centres on 25 January 2012. 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 Salzburg, Austria in February 2012. As had been the case 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 computer-based assessments were prepared and dispatched early in March 2012.

Main Survey coder query service

The coder query service operated for the Main Survey across all 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 could be removed from the international dataset, or in some cases from particular national datasets where an isolated problem occurred. One mathematics item was removed from the international data set as a result of this analysis. Further details on the outcomes of the analysis of Main Survey item data are provided in Chapter 12.

Released items

Several PISA items were released into the public domain at the time of publication of the PISA 2012 results, to illustrate the kinds of items used in the PISA assessment. Two intact clusters from the paper-based mathematics component of the

Main Survey, comprising 26 items, were released, along with a further 30 paper-based items that had been used in the Field Trial but were not selected for inclusion in the Main Survey item set. A further 11 mathematics items were released, from a cluster that had been used in the PISA 2006 administration but had subsequently been held in reserve. The items are available for download from the PISA website: *http://www.oecd.org/pisa/pisaproducts/*.

In addition, ten items from three units used in the computer-based mathematics component were released to supplement four units that had been put in the public domain prior to the assessment, along with four units of problem solving material to supplement the two that had been released earlier. Three additional reading units (to supplement the seven sample items previously posted) were added to the public website set up for this purpose. All of these computer-based items can be seen at *www.oecd.org/pisa*.

Some of these released paper-based items, including ten individual financial literacy items from the Field Trial that were not included in the Main Survey, were included in the publication *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy* (OECD, 2013), and some were used for illustrative purposes in the OECD international report of the *PISA 2012 Results* (OECD, 2014).

No new reading or science material was released after the 2012 survey administration.

Notes

1. One of those items was deleted internationally as a result of errors detected in the coding of responses.

2. Technical reference documents are available on the OECD PISA website: *www.oecd.org/pisa*.

References

Buchner, A. and **J. Funke** (1993), "Finite-state Automata: Dynamic Task Environments in Problem-solving Research", *The Quarterly Journal of Experimental Psychology*, Vol. 46A, pp. 83-118.

Funke, J. (2001), "Dynamic systems as tools for analysing human judgement", *Thinking and Reasoning*, Vol. 7, pp. 69-79.

Greiff, S., D. V. Holt, S. Wüstenberg, F. Goldhammer, and **J. Funke** (2013) "Computer-based assessment of complex problem solving: Concept, implementation, and application", *Educational Technology Research & Development*, Vol 61, pp. 407-421.

Postlethwaite and **A. Grisay** (eds.), *A view inside primary schools: A World Education Indicators (WEI) cross-national study*, Unesco Institute of Statistics, Montreal, pp. 175-207.

OECD (2014), *PISA 2012 Results: What Students Know and Can Do Student Performance in Mathematics, Reading and Science (Volume I Revised edition, February 2014),* PISA, OECD Publishing, Paris.
http://dx.doi.org/10.1787/9789264208780-en

OECD (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Sovling and Financial Literacy*, PISA, OECD Publishing, Paris.
http://dx.doi.org/10.1787/9789264190511-en

Ramalingam, D., B. McCrae and **R. Philpot** (2014), "The PISA Assessment of Problem Solving", Chapter 6, in Csapó, B. and J. Funke (eds.), *The Nature of Problem Solving*, OECD Publishing, Paris.

Wüstenberg, S., S. Greiff and **J. Funke** (2012), "Complex problem solving – More than reasoning?", *Intelligence*, Vol. 40, pp. 1-14.