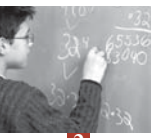


2

Test Design and Test Development



This chapter outlines the test design for PISA 2003, and describes the process by which the PISA consortium, led by ACER, developed the test instruments for use in PISA 2003.

TEST SCOPE AND FORMAT

In PISA 2003, four subject domains were tested, with mathematics as the major domain, and reading, science and problem solving as minor domains. Student achievement in mathematics was assessed using 85 test items representing approximately 210 minutes of testing time. This was a substantial reduction in the time allocated to the major domain for 2000 (reading), which had 270 minutes. The problem-solving assessment consisted of 19 items, the reading assessment consisted of 28 items and the science assessment consisted of 35 items, representing approximately 60 minutes of testing time for each of the minor domains.

The 167 items used in the main study were selected from a larger pool of approximately 300 items that had been tested in a field trial conducted by all national centres one year prior to the main study.

PISA 2003 was a paper-and-pencil test. The test items were multiple choice, short answer, and extended response. Multiple choice items were either standard multiple choice with a limited number (usually four) of 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 (true/false, correct/incorrect, etc.). Short answer items included both closed-constructed response items that generally required students to construct a response within very limited constraints, such as mathematics items requiring a numeric answer, and items requiring a word or short phrase, etc. Short-response items were similar to closed-constructed response items, but for these a wider range of responses was possible. Open-constructed response items required more extensive writing, or showing a calculation, and frequently included some explanation or justification. Pencils, erasers, rulers, and in some cases calculators, were provided. The consortium 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 items in the pool required a calculator, but some items involved solution steps for which the use of a calculator could facilitate computation. In developing the mathematics items, test developers were particularly mindful to ensure that the items were as calculator-neutral as possible.

TEST DESIGN

The 167 main study items were allocated to 13 item clusters (seven mathematics clusters and two clusters in each of the other domains), with each cluster representing 30 minutes of test time. The items were presented to students in 13 test booklets, with each booklet being composed of four clusters according to the rotation design shown in Table 2.1. M1 to M7 denote the mathematics clusters, R1 and R2 denote the reading clusters, S1 and S2 denote the science clusters, and PS1 and PS2 denote the problem-solving clusters. Each cluster appears in each of the four possible positions within a booklet exactly once. Each test item, therefore, appeared in four of the test booklets. This linked design enabled standard measurement techniques to be applied to the resulting student response data to estimate item difficulties and student abilities.

The sampled students were randomly assigned one of the booklets, which meant each student undertook two hours of testing.



Table 2.1 ■ Cluster rotation design used to form test booklets for PISA 2003

Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	M1	M2	M4	R1
2	M2	M3	M5	R2
3	M3	M4	M6	PS1
4	M4	M5	M7	PS2
5	M5	M6	S1	M1
6	M6	M7	S2	M2
7	M7	S1	R1	M3
8	S1	S2	R2	M4
9	S2	R1	PS1	M5
10	R1	R2	PS2	M6
11	R2	PS1	M1	M7
12	PS1	PS2	M2	S1
13	PS2	M1	M3	S2

In addition to the 13 two-hour booklets, a special one-hour booklet, referred to as the UH booklet (or the Une Heure booklet) was prepared for use in schools catering exclusively to students with special needs. The UH booklet was shorter, and contained items deemed most suitable for students with special educational needs. The UH booklet contained seven mathematics items, six reading items, eight science items and five problem-solving items.

The two-hour test booklets were arranged in two one-hour parts, each made up of two of the 30-minute time blocks from the columns in the above figure. PISA's procedures provided for a short break to be taken between administration of the two parts of the test booklet, and a longer break to be taken between administration of the test and the questionnaire.

DEVELOPMENT TIMELINE

Detailed consortium planning of the development of items for PISA 2003 commenced in March 2000. Initial planning documents addressed the following key issues:

- Potential contributors to the development of items in the various domains;
- The need to ensure that the frameworks were sufficiently developed to define the scope and nature of items required for each domain, particularly in mathematics and problem solving;
- The various cognitive laboratory procedures that would be implemented; and
- The major development steps and timeline for the development process.

The PISA 2003 project started formally in September 2000, and concluded in December 2004. Among the first tasks for the project was establishing the relevant expert committees, including the mathematics expert group, to revise and expand the framework that had been used for the PISA 2000 assessment. A problem-solving expert group was also established to develop a framework for that part of the assessment. A major purpose of those frameworks was to define the test domain in sufficient detail to permit test development to proceed. The formal process of test development began after the first SMEG meetings in February 2001, although preliminary item development work started in September 2000. The main



phase of the test item development finished when the items were distributed for the field trial in December 2001. During this ten-month period, intensive work was carried out in writing and reviewing items, and in conducting cognitive laboratory activities. The field trial for most countries took place between February and July 2002, after which items were selected for the main study and distributed to countries in December 2002. Table 2.2 shows the major milestones and activities of the PISA 2003 test development timeline.

Table 2.2 ■ Test development timeline

Activity	Period
Develop frameworks	September 2000 - July 2001
Develop items	September 2000 - October 2001
Item submission from countries	February - June 2001
National item reviews	February - October 2001
Distribution of field trial material	November - December 2001
Translation into national languages	December 2001 - February 2002
Field trial coder training	February 2002
Field trial in participating countries	February - July 2002
Select items for main study	July - October 2002
Preparation of final source versions of all main study materials, in English and French	October - December 2002
Distribute main study material	December 2002
Main study coder training	February 2003
Main study in participating countries	February - October 2003

TEST DEVELOPMENT PROCESS

The test development process commenced with preparation of the assessment frameworks, review and refinement of test development methodologies and training of the relevant personnel in those methodologies. The process continued with calling for 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, preparing coding guides and coder training material, and selecting and preparing items for the main study.

Development of the assessment frameworks

The first major development task was to produce a set of assessment frameworks for each cognitive domain of the PISA assessment in accordance with the policy requirements of the PGB. The consortium, through the test developers and expert groups, and in consultation with national centres, and with regular consultation with national experts through the Mathematics Forum, developed a revised and expanded assessment framework for mathematics. A framework was developed using a similar process for problem solving. This took place in the latter part of 2000, and during 2001, with final revisions and preparation for publication during 2002. The frameworks were endorsed by the PISA Governing Board and published



in *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills* (OECD, 2003). The frameworks presented the direction being taken by the PISA assessments. They defined each assessment domain, described the scope of the assessment, the number of items required to assess each component of a domain and the preferred balance of question types, and sketched the possibilities for reporting results.

Development and documentation of procedures

The terms of reference for the PISA 2003 contract contained references to the use of cognitive laboratory procedures in the development of test items, including the following:

Different from the first survey cycle, the contractor shall also be expected to use new techniques and methods for the development of the item pool. For instance, cognitive laboratory testing of items may be useful in filtering out, even prior to the field test, poorly functioning items.

And later, in the project's terms of reference:

The contractor shall provide evidence from cognitive laboratories that student responses to items on the assessment are indeed reflective of the cognitive activities they were designed to sample. The contractor shall develop protocols for collecting input from students that reflects their approaches to the problems and which gives evidence about how they approached and solved the various problems. Without such information, interpretations of student response data may reflect a high level of inference.

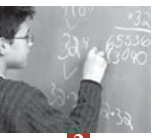
In response to this the consortium carried out research into practices employed under the title cognitive laboratories, and reviewed existing item development practices in light of that research. A methodology was developed that combined existing practices, together with refinements gleaned from the research literature on cognitive laboratories, which met the requirements of the terms of reference. The methodology included the following key elements:

- Cognitive walk-through (otherwise known as item panning, or item shredding);
- Cognitive interviews (including individual think-aloud methods involving the recording of individual students as they worked on items, cognitive interviews with individual students, and cognitive group interviews); and
- Cognitive comparison studies (including pre-pilot studies and other pilot testing of items with groups of students).

Test developers at the various consortium item development centres were briefed on the methodology, and the procedures were applied as far as possible in the development of all items. Cognitive walk-throughs were employed on all items developed, cognitive interviews were employed on a significant proportion of items, and cognitive comparison studies were used for all items.

Item submission guidelines

An international comparative study should ideally draw items from a wide range of cultural settings and languages. A comprehensive set of guidelines for the submission of mathematics items was developed and distributed to national project managers in February 2001 to encourage national submission of items from as many participating countries as possible. The item submission guidelines for mathematics are included in Appendix 4. Similar guidelines were also developed for the problem-solving domain. The



guidelines included an overview of the development process and timelines, as well as significant detail on the requirements for writing items, relationships with the mathematics framework, and a discussion of issues affecting item difficulty. A number of sample items were also provided. An item submission form accompanied the guidelines, to assist with identification and classification of item submissions. A final deadline for submission of items was set as the end of June 2001.

National item submissions

Item submissions in mathematics were received from 15 countries, between January and July 2001. Countries formally submitting items were Argentina, Austria, Canada, Czech Republic, Denmark, France, Germany, Ireland, Italy, Japan, Korea, Norway, Portugal, Sweden and Switzerland. Approximately 500 items were submitted, and items were submitted in seven different languages (English, French, German, Italian, Japanese, Portuguese and Spanish). The smallest submission was a single unit comprising three items. The largest was a collection of 60 units comprising about 106 items.

In addition to the three consortium centres involved in problem-solving item development (ACER in Australia, CITO in the Netherlands and a group at the University of Leeds in the United Kingdom), items were also submitted by the national centres of Italy, Ireland and Brazil. From the submitted material, seven units (comprising 40 items) were included in the material sent to all countries for review.

Some submitted items had already undergone significant development work, including field-testing with students, prior to submission. Others were in a much less developed state and consisted in some cases of little more than some stimulus material and ideas for possible questions. All submitted material required significant additional work by consortium test developers.

Development of test items

A complete PISA item consists of some stimulus material, one or more questions, and a guide to the coding of responses to each question. The coding guides comprise a list of response categories, each with its own scoring code, descriptions of the kinds of responses to be assigned each of the codes, and sample responses for each response category.

One other feature of test items that was developed for PISA 2000 and continued for PISA 2003 relates to double-digit coding, which can be used to indicate both the score and the response code. The double-digit codes allow distinctions to be retained between responses that are reflective of quite different cognitive processes and knowledge. For example, if an algebraic approach or a trial-and-error approach was used to arrive at a correct answer, a student could score a '1' for an item using either of these methods, and the method used would be reflected in the second digit. The double-digit coding captures different problem-solving approaches by using the first digit to indicate the score and the second digit to indicate method or approach.

The development of mathematics items took place at one or more of the consortium item development centres: The ACER in Australia, CITO in the Netherlands and NIER in Japan. Item development in problem solving was carried out at ACER, CITO and the University of Leeds. Professional item developers at each of the centres wrote and developed items. In addition, items received from national submissions or from individuals wishing to submit items (for example individual members of the mathematics expert group also submitted a number of items for consideration) were distributed among the relevant item development centres for the required development work.



Typically, the following steps were followed in the development of items, including both items originating at the consortium centre concerned and items from national submissions that were allocated to each consortium centre for development. The steps are described in a linear fashion, but in reality they were often negotiated in a cyclic fashion, with items typically going through the various steps more than once. The steps were:

Initial preparation

A professional item writer prepared items in a standard format, including item stimulus, one or more questions, and a proposed coding guide for each question.

Item panelling

Each item was given extensive scrutiny at a meeting of a number of professional item writers. This stage of the cognitive laboratory process typically involved item writers in a vigorous analysis of all aspects of the item, including from the point of view both students and coders.

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

Cognitive interview

Many items were then prepared for individual students or small groups of students to attempt. A combination of think-aloud methods, individual interviews and group interviews were used with students to ascertain the thought processes typically employed by students as they attempt the items.

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

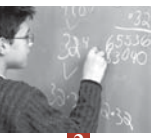
International item panelling

All items were scrutinised by panels of professional item writers in at least two of the item development centres. The feedback provided, following scrutiny of items by international colleagues, assisted the item development teams to introduce further improvements to the items.

Pilot testing

Every item that was developed was subjected to pilot testing in schools with a substantial number of students who were in the relevant age range. Test booklets were formed from a number of items. These booklets were field tested with several whole classes of students in several different schools. Piloting of this kind took place in schools in Australia, Japan, the Netherlands and Austria. Frequently, multiple versions of items were field tested, and the results were compared to ensure that the best alternative form was identified. Data from the field testing were analysed using standard item response techniques.

Items were revised, often extensively, following pilot testing with large groups of students. In some cases, revised versions of items were again subjected to the pilot testing procedure. One of the most important outputs of this stage of the cognitive laboratory procedures was the generation of student responses to all questions. A selection of these responses were added to the scoring guides to provide additional sample answers, showing coders how to code a variety of different responses to each item.



At the conclusion of these steps, surviving items were considered ready for circulation to national centres for review and feedback.

NATIONAL REVIEW OF ITEMS

In February 2001, National Project Managers were given a set of item review guidelines to assist them in reviewing items and providing feedback. A copy of a similar set of guidelines, prepared later for review of all items used in the field trial, is appended to this document (see Appendix 5). A central aspect of that review was a request to national experts to rate items according to various features, including their relevance and acceptability from a cultural perspective. Specific issues and problems that might be associated with cultural differences among countries were also identified at that time. Other features on which national experts commented were interest, curriculum relevance, relevance to the PISA framework, and any other matters thought to be important by any national centre.

NPMs were also given a schedule for the distribution and review of draft items that would occur during the remainder of 2001.

As items were developed to a sufficiently complete stage, they were dispatched to national centres for review. Four bundles of items were sent. The first bundle, comprising 106 mathematics items, was dispatched on 30 March 2001. National centres were given a feedback form, which drew attention to various matters of importance for each item, and were asked to provide detailed feedback within four weeks. Subsequent bundles were dispatched on 3 May (comprising 29 problem-solving items), 3 June (comprising 28 problem-solving items and 179 mathematics items) and 7 August (comprising 45 problem-solving items, 115 mathematics items and 38 science items). In each case, NPMs were given four weeks to gather feedback from the relevant national experts, and return the completed feedback forms to the consortium.

The feedback from NPMs was collated into a small set of reports, and made available to all NPMs on the PISA Web site. The reports were used extensively at meetings of the mathematics forum and the mathematics, problem-solving and science expert groups as they considered the items being developed. The feedback frequently resulted in further significant revision of the items. In particular, issues related to translation of items into different languages were highlighted at this stage, as were other cultural issues related to the potential operation of items in different national contexts.

INTERNATIONAL ITEM REVIEW

As well as this formal, structured process for national review of items, the bundles of mathematics items were also considered in detail at meetings of the mathematics forum. All PISA countries were invited to send national mathematics experts to meetings of the forum. At the meeting that took place in Lisbon, Portugal, in May 2001, all items that had been developed at that stage were reviewed in detail. Significant feedback was provided, resulting in revisions to many of the items.

A similar review process involving the mathematics expert group was also employed. Meetings of the group in February, July and September 2001 spent considerable time reviewing mathematics items in great detail. Problem-solving and science items were similarly scrutinised by the relevant expert groups.



A further small bundle of late developed or significantly revised mathematics items was prepared, and reviewed by the mathematics forum¹ and the mathematics expert group at a joint meeting held in Nijmegen, the Netherlands, in September 2001.

FRENCH TRANSLATION

When items reached the stage of readiness for national review, they were also considered to be ready for translation into French. At that time they were entered in a web-based item-tracking database. Test developers and consortium translation staff used this facility to track the parallel development of English and French language versions.

Part of the translation process involved verification by French subject experts, who were able to identify issues 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 verification process, which assisted in ensuring that items were as culturally neutral as possible, in identifying instances of wording that could be modified to simplify translation into other languages, and in identifying particular items where translation notes were needed to ensure the required accuracy in translating items to other languages.

ITEM POOL

A total of 512 mathematics items were developed to the stage where they were suitable for circulation to national centres for feedback, and could be seriously considered for inclusion in the test instruments for the PISA 2003 study. A further 20 items were retained from PISA 2000 for possible use as link items. Similarly, a total of 102 new problem-solving items and 38 new science items were developed to this stage, and circulated to national centres for review.

FIELD TRIAL ITEMS

In September 2001 the items to be used in the 2002 field trial were selected from the item pool. A joint meeting of the mathematics forum and the mathematics expert group was held in Nijmegen, the Netherlands, in September 2001 to commence the selection process. Participants rated items, and assigned each item a priority for inclusion in the field trial pool. A number of items were identified for rejection from the pool.

The MEG continued the selection task over the two days following, and presented a set of 237 recommended items to a meeting of NPMs the following week. The problem-solving and science expert groups also selected items for the problem-solving and science instruments, and presented these to the same NPM meeting.

The consortium carefully considered the advice from the national item feedback, the mathematics forum, the three expert groups, and the NPM meeting. Consortium item developers made further refinements to the selection of recommended items where necessary for purposes of balance in relation to framework requirements, and the consortium selected a final set of items for the field trial. A total of 217 mathematics items, 35 science items and 51 problem-solving items were selected. Some of the important characteristics of the selected mathematics items are summarised in Table 2.3, Table 2.4 and Table 2.5.

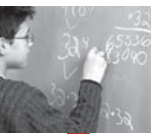


Table 2.3 ■ Mathematics field trial items (item format by competency cluster)

Item format	Competency cluster			
	Reproduction	Connections	Reflection	Total
Multiple-choice response	13	44	22	79
Closed-constructed response	28	31	10	69
Open-constructed response	10	37	22	69
Total	51	112	54	217

Table 2.4 ■ Mathematics field trial items (content category by competency cluster)

Content category	Competency cluster			
	Reproduction	Connections	Reflection	Total
Space and shape	12	20	7	39
Quantity	19	30	9	58
Change and relationships	11	38	21	70
Uncertainty	9	24	17	50
Total	51	112	54	217

Table 2.5 ■ Mathematics field trial items (content category by item format)

Content category	Item format			Total
	Multiple-choice response	Closed-constructed response	Open-constructed response	
Space and shape	11	12	16	39
Quantity	17	26	15	58
Change and relationships	30	18	22	70
Uncertainty	21	13	16	50
Total	79	69	69	217



The important framework characteristics of the problem-solving and science items are summarised in Table 2.6 and Table 2.7.

Table 2.6 ■ Problem-solving field trial items (problem type by item format)

Problem-solving type	Item format			Total
	Closed-constructed response	Multiple-choice response	Open-constructed response	
Decision making	2	6	12	20
System analysis and design	1	10	8	19
Trouble shooting	0	9	3	12
Total	3	25	23	51

Table 2.7 ■ Science field trial items (science process by item format)

Science process	Item format					Total
	Closed-constructed response	Complex multiple-choice response	Multiple-choice response	Open-constructed response	Short response	
Describing, explaining and predicting	1	6	4	5	2	18
Interpreting scientific evidence	0	1	5	8	0	14
Understanding scientific investigation	0	0	3	0	0	3
Total	1	7	12	13	2	35

The mathematics items were placed into 14 clusters, each designed to represent 30 minutes of testing. Likewise, four clusters of problem-solving items and two clusters of science items were formed. The clusters were then placed into ten test booklets according to the field trial test design, shown in Table 2.8. Each booklet contained four clusters.

Table 2.8 ■ Allocation of item clusters to test booklets for field trial

Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	M1	M11	S2	M2
2	M2	M12	M11	M3
3	M3	M13	M12	M4
4	M4	M14	M13	M5
5	M5	P1	M14	M6
6	M6	P2	P1	M7
7	M7	P3	P2	M8
8	M8	P4	P3	M9
9	M9	S1	P4	M10
10	M10	S2	S1	M1



The final forms of all selected items were subjected to a final editorial check using the services of a professional editor. This assisted in uncovering remaining grammatical inconsistencies and other textual and layout irregularities, and ensuring high quality in the presentation of the final product.

English and French versions of items, clusters and booklets were distributed to national centres in three dispatches, on 1 November, 16 November and 3 December 2001. A consolidated dispatch of all items, clusters and booklets, including errata, as well as other material for the field trial, was sent on compact disk to all countries on 21 December.

National centres then commenced the process of preparing national versions of all selected items. 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 response coders. Coder training sessions for mathematics, problem solving, reading and science were scheduled for February 2002. Consolidated coding guides were prepared, in both English and French, containing all those items that required manual coding. The guide 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. The actual scoring was done after the field trial data were analysed, as the analysis was used to provide information on the appropriate scores for each different response category². The Coding Guide was a list of response codes with descriptions and examples, but a separate training workshop document was also produced for each subject area, which consisted of additional student responses to the items, which could be used for practice coding and discussion at the coder training sessions.

All countries sent representatives to the training sessions, which were conducted in Salzburg, Austria, in February 2002. As a result of the use of the coding guides in the training sessions, the need to introduce a small number of further amendments to coding guides was identified. These amendments were incorporated in a final dispatch of coding guides and training materials, on 14 March 2002, after the Salzburg training meetings. Following the training sessions, national centres recruited coders, and conducted their own training in preparation for the coding of field trial scripts.

FIELD TRIAL CODER QUERIES

The consortium provided a coder query service to support NPMs running the coding of scripts in each country. When there was any uncertainty, national centres were able to submit queries by telephone or email 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 and consortium responses to those queries were published on the consortium website. The queries report was regularly updated as new queries were received and dealt with. This meant that all national coding centres had access to an additional source of advice about responses that had been found at all problematic. Coding supervisors in all countries found this to be a particularly useful resource.



FIELD TRIAL OUTCOMES

Extensive analyses were conducted on the field trial item response data. These analyses included the standard *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 (see Chapter 9).

On the basis of these critical measurement characteristics, a proportion of the field trial items were identified as having failed the trial and were marked for removal from the pool of items that would be considered for the main study.

A timing study was conducted to gather data on the average time taken to respond to items. A multiple coder study was carried out to investigate the inter-coder reliability of manually coded items.

NATIONAL REVIEW OF FIELD TRIAL ITEMS

In addition, a further round of national rating of items was carried out, with a view to gaining ratings of field trial items informed by the experience at national centres of how the items actually worked in each country. A set of review guidelines was designed to assist national experts to focus on the most important features of possible concern (Appendix 5). Almost all countries submitted this final set of priority ratings of all field trial items for possible inclusion in the main study item pool.

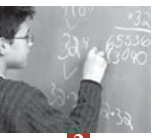
Further, a comprehensive field trial review report was prepared by all NPMs. These reports included a further opportunity for NPMs to identify particular strengths and weaknesses of individual items, stemming from the translation and verification process, preparation of test forms, coding of student responses to items, and entry of data.

MAIN STUDY ITEM SELECTION

Subject matter expert groups for mathematics, science, problem solving and reading met in October 2002 in Melbourne, Australia, to review all available material and formulate recommendations about items to be included in the main study item pool. They took into account all available information, including the field trial data, national item rating data, information coming from the translation process, information from the national field trial reviews, and the constraints imposed by the assessment framework for each domain.

For the mathematics domain, a total of 65 items were needed from the field trial pool of 217. The selection had to satisfy the following constraints:

- The number of items (about 65) was based on the results of the timing study, which concluded that thirty-minute item clusters should contain an average of 12 to 13 items;
- The major framework categories (overarching ideas, and competency clusters) had to be populated according to the specifications of the framework;
- The proportion of items that required manual coding had to be limited to around 40 per cent;
- The psychometric properties of all selected items had to be satisfactory;
- Items that generated coding problems were to be avoided unless those problems could be properly addressed through modifications to the coding instructions;



- Items given high priority ratings by national centres were preferred, and items with lower ratings were to be avoided; and
- Once all these characteristics were satisfied, items reflecting mathematical literacy in an interesting way would be preferred.

The mathematics expert group identified a total of 88 items suitable for possible inclusion in the main study, including the 20 items retained for linking purposes from the PISA 2000 test. The science expert group identified 10 new items to replace the 10 that had been released from the PISA 2000 item set. This meant they had a set of 37 items recommended for inclusion in the PISA 2003 main study. The problem-solving expert group identified 20 items suitable for inclusion. The reading expert group recommended a selection of 33 items from the PISA 2000 main study item pool for inclusion in the PISA 2003 instruments.

The consortium carefully considered the advice from the four expert groups, and made some adjustments to the recommended selections in reading (by removing four items, reducing the final pool to 29 items) and in mathematics. The adjustments to the mathematics selection were a little more extensive in order to resolve a number of remaining problems with the initial preferred selection of the expert group:

- The total number of items selected had to be reduced from 88 to a maximum of 85;
- The overall difficulty of the selection had to be reduced;
- The number of relatively easy items had to be increased slightly; and
- A small number of items that had relatively high omission rates had to be removed from the selection.

These adjustments had to be made while retaining the required balance of framework categories. In the end a total of 85 mathematics items were selected (including 20 that were retained for linking purposes from the PISA 2000 study). The final selection included a small number of items that had been given relatively low ratings by national centres. These items were needed either to reduce average item difficulty, or because they were seen to contribute something important to the way the test reflected the framework conception of mathematical literacy. Similarly, a number of items that had been highly rated were not included. These items suffered from one of more problems, including poor psychometric properties, being too difficult, or there were remaining problems with use of the coding guides.

The proposed selection was presented to the PGB in Prague, Czech Republic in October 2002, and to a meeting of National Project Managers in Melbourne also in October. The characteristics of the final item selection, with respect to the major framework categories, are summarised in Table 2.9, Table 2.10 and Table 2.11.

Table 2.9 ■ Mathematics main study items (item format by competency cluster)

Item format	Competency cluster			Total
	Reproduction	Connections	Reflection	
Multiple-choice response	7	14	7	28
Closed-constructed response	7	4	2	13
Open-constructed response	12	22	10	44
Total	26	40	19	85



Table 2.10 ■ Mathematics main study items (content category by competency cluster)

Content category	Competency cluster			Total
	Reproduction	Connections	Reflection	
Space and shape	5	12	3	20
Quantity	9	11	3	23
Change and relationships	7	8	7	22
Uncertainty	5	9	6	20
Total	26 (31%)	40 (47%)	19 (22%)	85

Table 2.11 ■ Mathematics main study items (content category by item format)

Content category	Item format			Total
	Multiple-choice response	Closed-constructed response	Open-constructed response	
Space and shape	8	6	6	20
Quantity	6	2	15	23
Change and relationships	3	4	15	22
Uncertainty	11	1	8	20
Total	28	13	44	85

For the reading domain, 28 items were selected from the PISA 2000 item pool for use in the PISA 2003 main study. Items were selected from the PISA 2000 items with the best psychometric characteristics, and to retain a balance in the major framework categories. Some of the framework characteristics of the selected items are summarised in Table 2.12 and Table 2.13.

For the problem-solving domain, 19 items were selected for use in the main study. Their major characteristics are summarised in Table 2.14.

For the science domain, 35 items were selected, including 20 that had been retained from the PISA 2000 main study item pool, and 15 new items that had been selected from those items used in the field trial. Their major characteristics are summarised in Table 2.15.

Table 2.12 ■ Reading main study items (reading process by item format)

Reading process	Item format			Short response	Total
	Closed-constructed response	Multiple-choice response	Open-constructed response		
Retrieving information	3	1	0	3	7
Interpreting	1	9	3	1	14
Reflecting	0	0	7	0	7
Total	4	10	10	4	28

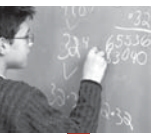


Table 2.13 ■ Reading main study items (text structure type by item format)

Text structure type	Item format				Total
	Closed-constructed response	Multiple-choice response	Open-constructed response	Short response	
Continuous	0	9	9	0	18
Non-continuous	4	1	1	4	10
Total	4	10	10	4	28

Table 2.14 ■ Problem solving main study items (problem type by item format)

Problem-solving type	Item format			Total
	Closed-constructed response	Multiple-choice response	Open-constructed response	
Decision making	2	2	3	7
System analysis and design	1	2	4	7
Trouble shooting	0	3	2	5
Total	3	7	9	19

Table 2.15 ■ Science main study items (science process by item format)

Science process	Item format				Total
	Complex-multiple choice	Multiple-choice response	Open-constructed response	Short response	
Describing, explaining and predicting	3	7	6	1	17
Interpreting scientific evidence	2	4	5	0	11
Understanding scientific investigation	2	2	3	0	7
Total	7	13	14	1	35

After finalising the main study 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. A further round of professional editing took place. French translations of all selected items were updated. Clusters of items were formed in each of the four test domains in accordance with the main study rotation design, shown previously in Table 2.1. Test booklets were prepared in English and French.

All items, item clusters and test booklets, in English and French, were dispatched to national centres in three dispatches, on 10 December, 13 December and 20 December 2002.

This enabled national centres to finalise the required revisions to their own national versions of all selected test items, and to prepare test booklets for the main study.



MAIN STUDY CODER TRAINING

Following final selection and dispatch of items to be included in the main study, various documents and materials were prepared to assist in the training of coders. Coder training sessions for mathematics, problem solving, reading and science were scheduled for February 2003. Consolidated coding guides were prepared, in both English and French, containing all those items that required manual coding. These were dispatched to national centres in early January 2003. In addition, the training materials prepared for the field trial coder training were revised and expanded, with additional student responses to the items. These additional responses were gathered during the field trial and in particular from the coder query service that had operated during the field trial coding. They were chosen for use in practice coding and discussion at the coder training sessions.

Coder training sessions were conducted in Madrid, Spain, in February 2003. All but three countries had representatives at the training meetings. Arrangements were put in place to ensure appropriate training of representatives from those countries not in attendance.

Once again, a small number of clarifications were needed to make the coding guides and training materials as clear as possible, and revised coding guides and coder training materials were prepared and dispatched in March 2003 following the training meetings.

MAIN STUDY CODER QUERY SERVICE

The coder query service operated for the main study across the four test domains. Any student responses that national centre coders found difficult to code were 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 these reports were regularly updated as new queries were received.

REVIEW OF MAIN STUDY ITEM ANALYSES

On receipt of data from the main study testing, extensive analyses of item responses were carried out to identify any items that were not capable of generating useful student achievement data. Such items were identified for removal from the international dataset, or in some cases from particular national datasets where an isolated problem occurred.

Notes

- 1 The mathematics forum was a gathering of country representatives, nominated by PGB members, which had expertise in mathematics education and assessment.
- 2 It is worth mentioning here that as data entry was carried out using *KeyQuest*, many short responses were entered directly, which saved time and made it possible to capture students' raw responses.



READER'S GUIDE

Country codes

The following country codes are used in this report:

OECD countries

AUS	Australia
AUT	Austria
BEL	Belgium
BEF	Belgium (French Community)
BEN	Belgium (Flemish Community)
CAN	Canada
CAE	Canada (English Community)
CAF	Canada (French Community)
CZE	Czech Republic
DNK	Denmark
FIN	Finland
FRA	France
DEU	Germany
GRC	Greece
HUN	Hungary
ISL	Iceland
IRL	Ireland
ITA	Italy
JPN	Japan
KOR	Korea
LUX	Luxembourg
LXF	Luxembourg (French Community)
LXG	Luxembourg (German Community)
MEX	Mexico
NLD	Netherlands
NZL	New Zealand
NOR	Norway
POL	Poland
PRT	Portugal

SVK	Slovak Republic
ESP	Spain
ESB	Spain (Basque Community)
ESC	Spain (Catalonian Community)
ESS	Spain (Castillian Community)
SWE	Sweden
CHE	Switzerland
CHF	Switzerland (French Community)
CHG	Switzerland (German Community)
CHI	Switzerland (Italian Community)
TUR	Turkey
GBR	United Kingdom
IRL	Ireland
SCO	Scotland
USA	United States

Partner countries

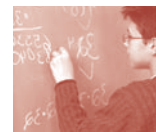
BRA	Brazil
HKG	Hong Kong-China
IND	Indonesia
LVA	Latvia
LVL	Latvia (Latvian Community)
LVR	Latvia (Russian Community)
LIE	Liechtenstein
MAC	Macao-China
RUS	Russian Federation
YUG	Serbia and Montenegro (Serbia)
THA	Thailand
TUN	Tunisia
URY	Uruguay



List of abbreviations

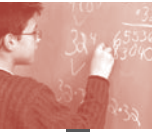
The following abbreviations are used in this report:

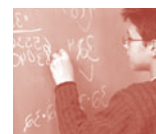
ACER	Australian Council for Educational Research	NDP	National Desired Population
AGFI	Adjusted Goodness-of-Fit Index	NEP	National Enrolled Population
BRR	Balanced Repeated Replication	NFI	Normed Fit Index
CFA	Confirmatory Factor Analysis	NIER	National Institute for Educational Research, Japan
CFI	Comparative Fit Index	NNFI	Non-Normed Fit Index
CITO	National Institute for Educational Measurement, The Netherlands	NPM	National Project Manager
CIVED	Civic Education Study	OECD	Organisation for Economic Cooperation and Development
DIF	Differential Item Functioning	PISA	Programme for International Student Assessment
ESCS	Economic, Social and Cultural Status	PPS	Probability Proportional to Size
ENR	Enrolment of 15-year-olds	PGB	PISA Governing Board
ETS	Educational Testing Service	PQM	PISA Quality Monitor
IAEP	International Assessment of Educational Progress	PSU	Primary Sampling Units
I	Sampling Interval	QAS	Questionnaire Adaptations Spreadsheet
ICR	Inter-Country Coder Reliability Study	RMSEA	Root Mean Square Error of Approximation
ICT	Information Communication Technology	RN	Random Number
IEA	International Association for the Evaluation of Educational Achievement	SC	School Co-ordinator
INES	OECD Indicators of Education Systems	SD	Standard Deviation
IRT	Item Response Theory	SEM	Structural Equation Modelling
ISCED	International Standard Classification of Education	SMEG	Subject Matter Expert Group
ISCO	International Standard Classification of Occupations	SPT	Study Programme Table
ISEI	International Socio-Economic Index	TA	Test Administrator
MENR	Enrolment for moderately small school	TAG	Technical Advisory Group
MOS	Measure of size	TCS	Target Cluster Size
NCQM	National Centre Quality Monitor	TIMSS	Third International Mathematics and Science Study
		TIMSS-R	Third International Mathematics and Science Study – Repeat
		VENR	Enrolment for very small schools
		WLE	Weighted Likelihood Estimates



References

- Adams, R.J., Wilson, M.R. and W. Wang** (1997), “The multidimensional random coefficients multinomial logit model”, *Applied Psychological Measurement* 21, pp. 1-24.
- Aiken, L. R.** (1974), “Two scales of attitudes toward mathematics,” *Journal for Research in Mathematics Education* 5, National Council of Teachers of Mathematics, Reston, pp. 67-71.
- Andersen, Erling B.** (1997), “The Rating Scale Model”, in van der Linden, W. J. and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory*, Springer, New York/Berlin/Heidelberg.
- Bandura, A.** (1986), *Social Foundations of Thought and Action: A Social Cognitive Theory*, Prentice Hall, Englewood Cliffs, N.J.
- Baumert, J. and O. Köller** (1998), “Interest Research in Secondary Level I : An Overview”, in L. Hoffmann, A. Krapp, K.A. Renninger & J. Baumert (eds.), *Interest and Learning*, IPN, Kiel.
- Beaton, A.E.** (1987), *Implementing the New Design: The NAEP 1983-84 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, N.J.
- Bryk, A. S. and S.W. Raudenbush** (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods*, SAGE Publications, Newbury Park.
- Bollen, K.A. and S.J. Long** (eds.) (1993), *Testing Structural Equation Models*, SAGE publications, Newbury Park.
- Branden, N.** (1994), *Six Pillars of Self-Esteem*. Bantam, New York.
- Brennan, R.L.** (1992), *Elements of Generalizability Theory*, American College Testing Program, Iowa City.
- Buchmann, C.** (2000), *Measuring Family Background in International Studies of Educational Achievement: Conceptual Issues and Methodological Challenges*, paper presented at a symposium convened by the Board on International Comparative Studies in Education of the National Academy of Sciences/National Research Council on 1 November, in Washington, D.C.
- Cochran, W.G.** (1977), *Sampling Techniques* (3rd edition), Wiley, New York.
- Cronbach, L.J., G.C. Gleser, H. Nanda and N. Rajaratnam** (1972), *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*, Wiley and Sons, New York.
- Eccles, J.S.** (1994), “Understanding Women’s Educational and Occupational choice: Applying the Eccles *et al.* Model of Achievement-Related Choices”, *Psychology of Women Quarterly* 18, Society for the Psychology of Women, Washington, D.C., pp. 585-609.

- 
- Eccles, J.S.** and **A. Wigfield** (1995), "In the mind of the achiever: The structure of adolescents' academic achievement-related beliefs and self-perceptions", *Personality and Social Psychology Bulletin* 21, Sage Publications, Thousand Oaks, pp. 215-225.
- Ganzeboom, H.B.G., P.M. de Graaf** and **D.J. Treiman** (1992), "A standard international socio-economic index of occupational status", *Social Science Research* 21, Elsevier, pp. 1-56.
- Gifi, A.** (1990), *Nonlinear Multivariate Analysis*, Wiley, New York.
- Greenacre, M.J.** (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, London.
- Grisay, A.** (2003), "Translation procedures in OECD/PISA 2000 international assessment", *Language Testing* 20, Holder Arnold Journals, pp. 225-240.
- Gustafsson, J.E** and **P.A. Stahl** (2000), *STREAMS User's Guide, Version 2.5 for Windows*, MultivariateWare, Mölndal, Sweden.
- Hacket, G.** and **N. Betz.** (1989), "An Exploration of the mathematics Efficacy/mathematics Performance Correspondence", *Journal of Research in Mathematics Education* 20, National Council of Teachers of Mathematics, Reston, pp. 261-273.
- Harvey-Beavis, A.** (2002), "Student and Questionnaire Development" in OECD, *PISA 2000 Technical Report*, OECD, Paris.
- Hatcher, L.** (1994), *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*, SAS Institute Inc., Cary.
- International Labour Organisation** (1990), *International Standard Classification of Occupations: ISCO-88*, International Labour Office, Geneva.
- Jöreskog, K.G.** and **Dag Sörbom** (1993), *LISREL 8 User's Reference Guide*, Scientific Software International, Chicago.
- Judkins, D.R.** (1990), "Fay's Method for Variance Estimation", *Journal of Official Statistics* 6, Statistics Sweden, Stockholm, pp. 223-239.
- Kaplan, D.** (2000), *Structural Equation Modeling: Foundation and Extensions*, SAGE Publications, Thousand Oaks.
- Keyfitz, N.** (1951), "Sampling with probabilities proportionate to science: Adjustment for changes in probabilities", *Journal of the American Statistical Association* 46, American Statistical Association, Alexandria, pp. 105-109.
- Lepper, M. R.** (1988), "Motivational considerations in the study of instruction", *Cognition and Instruction* 5, Lawrence Erlbaum Associates, Mahwah, pp. 289-309.
- Ma, X.** (1999), "A Meta-Analysis of the Relationship Between Anxiety Toward mathematics and Achievement in mathematics", *Journal for Research in Mathematics Education* 30, National Council of Teachers of Mathematics, Reston, pp. 520-540.



Macaskill, G., R.J. Adams and M.L. Wu (1998), “Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scales”, in M. Martin and D.L. Kelly (eds.) *Third International Mathematics and Science Study, Technical Report Volume 3: Implementation and Analysis*, Center for the Study of Testing, Evaluation and Educational Policy, Boston College, Chestnut Hill.

Marsh, H. W. (1990), *Self-Description Questionnaire (SDQ) II: A theoretical and Empirical Basis for the Measurement Of Multiple Dimensions of Adolescent Self-Concept: An Interim Test Manual and a Research Monograph*, The Psychological Corporation, San Antonio.

Marsh, H. W. (1994), “Confirmatory factor analysis models of factorial invariance: A multifaceted approach” *Structural Equation Modeling 1*, Lawrence Erlbaum Associates, Mahwah, pp. 5-34.

Marsh, H. W. (1999), *Evaluation of the Big-Two-Factor Theory of Motivation Orientation: Higher-order Factor Models and Age-related Changes*, paper presented at the 31.62 Symposium, Multiple Dimensions of Academic Self-Concept, Frames of Reference, Transitions, and International Perspectives: Studies From the SELF Research Centre. Sydney: University of Western Sydney.

Masters, G. N. and B. D. Wright (1997), “The Partial Credit Model”, in W. J. van der Linden and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory*, Springer, New York/Berlin/Heidelberg.

Meece, J., A. Wigfield and J. Eccles (1990), “Predictors of Maths Anxiety and its Influence on Young Adolescents’ Course Enrolment and Performance in Mathematics”, *Journal of Educational Psychology 82*, American Psychological Association, Washington, D.C., pp. 60-70.

Middleton, J.A. and P.A. Spanias (1999), “Findings, Generalizations, and Criticisms of the Research”, *Journal for Research in Mathematics Education 30*, National Council of Teachers of Mathematics, Reston, pp. 65-88.

Mislevy, R.J. (1991), “Randomization-based inference about latent variable from complex samples”, *Psychometrika 56*, Psychometric Society, Greensboro, pp. 177-196.

Mislevy, R.J. and K.M. Sheehan (1987), “Marginal estimation procedures”, in A.E. Beaton (ed.), *The NAEP 1983-1984 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, N.J.

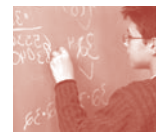
Mislevy, R.J. and K.M. Sheehan (1980), “Information matrices in latent-variable models”, *Journal of Educational Statistics 14.4*, American Educational Research Association and American Statistical Association, Washington, D.C., and Alexandria, pp. 335-350.

Mislevy, R.J., A.E. Beaton, B. Kaplan and K.M. Sheehan. (1992), “Estimating population characteristics form sparse matrix samples of item responses”, *Journal of Educational Measurement 29*, National Council on Measurement in Education, Washington, D.C., pp. 133-161.

Multon, K. D., S. D. Brown and R.W. Lent (1991), “Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation”, *Journal of Counselling Psychology 38*, American Psychological Association, Washington, D.C., pp. 30-38.

Muthén, B. O., S. H. C. du Toit and D. Spisic (1997), “Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical outcomes”, *Psychometrika*, Psychometric Society, Greensboro.

- 
- Muthen, L. and B. Muthen** (2003), *Mplus User's Guide Version 3.1*, Muthen & Muthen, Los Angeles.
- Nishisato, S.** (1980), *Analysis of Categorical Data: Dual Scaling and its Applications*, University of Toronto Press, Toronto.
- OECD** (Organisation for Economic Co-Operation and Development) (1999), *Classifying Educational Programmes: Manual for ISCED-97 Implementation in OECD Countries*, OECD, Paris.
- OECD** (2001), *Knowledge and Skills for Life: First Results from PISA 2000*, OECD, Paris.
- OECD** (2002), *PISA 2000 Technical Report*, OECD, Paris.
- OECD** (2003), *Student Engagement at School: A Sense of Belonging and Participation: Results from PISA 2000*, OECD, Paris.
- OECD** (2004a), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, OECD, Paris.
- OECD** (2004b), *Learning for Tomorrow's World – First Results from PISA 2003*, OECD, Paris.
- OECD** (2004c), *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003*, OECD, Paris.
- OECD** (2005a), *PISA 2003 Data Analysis Manual: SAS[®] Users*, OECD, Paris.
- OECD** (2005b), *PISA 2003 Data Analysis Manual: SPSS[®] Users*, OECD, Paris.
- Owens L. and J. Barnes** (1992), *Learning Preference Scales*, Australian Council for Educational Research, Hawthorn.
- Rasch, G.** (1960), *Probabilistic models for some intelligence and attainment tests*, Nielsen and Lydiche, Copenhagen.
- Rust, K.** (1985), "Variance estimation for complex estimators in sample surveys", *Journal of Official Statistics 1*, Statistics Sweden, Stockholm, pp. 381-397.
- Rust, K. and J.N.K. Rao** (1996), "Variance estimation for complex surveys using replication techniques", *Statistical Methods in Medical Research 5*, Holder Arnold Journals, pp. 283-310.
- Sändal, C.E., B. Swensson and J. Wretman** (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Schaffer, E. C., P.S. Nesselrodt and S. Stringfield** (1994), "The Contribution of Classroom Observation to School Effectiveness Research" in Reynolds *et. al.* (eds.), *Advances in School Effectiveness Research and Practice*, Pergamon, Oxford/New York/Tokyo.
- Schulz, W.** (2003), *Validating Questionnaire Constructs in International Studies. Two Examples from PISA 2000*, paper presented at the Annual Meeting of the American Educational Research Association (AERA) in Chicago, 21-25 April.

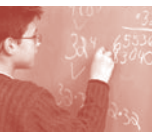


- Schulz, W.** (2004), "Mapping Student Scores to Item Responses", in W. Schulz and H. Sibberns (eds.), *IEA Civic Education Study. Technical Report*, IEA, Amsterdam.
- Sirotnik, K.** (1970), "An analysis of variance framework for matrix sampling", *Educational and Psychological Measurement* 30, SAGE Publications, pp. 891-908.
- Slavin, R. E.** (1983), "When does cooperative learning increase student achievement?" *Psychological Bulletin* 94, American Psychological Association, Washington, D.C., pp. 429-445.
- Statistical Solutions** (1992), *BMDP Statistical Software*, Statistical Solutions, Los Angeles.
- Teddlie, C. and D. Reynolds** (2000) (eds.), *The International Handbook of School Effectiveness Research*, Falmer Press, London/New York.
- Thorndike, R.L.** (1973), *Reading Comprehension Education in Fifteen Countries: An Empirical Study*, Almquist & Wiksell, Stockholm.
- Travers, K. J., R.A. Garden and M. Rosier** (1989), "Introduction to the Study", in D.A. Robitaille and R.A. Garden (eds.), *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics Curricula*, Pergamon Press, Oxford.
- Travers, K. J. and I. Westbury** (1989), *The IEA Study of Mathematics I: Analysis of Mathematics Curricula*, Pergamon Press, Oxford.
- Verhelst, N.** (2004), "Generalizability Theory", in Council of Europe, *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, (Section E), Council of Europe (DGIV/EDU/LANG (2004) 13), Strasbourg.
- Warm, T. A.** (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika* 54, Psychometric Society, Greensboro, pp. 427-45.
- Wigfield, A., J. S. Eccles and D. Rodriguez** (1998), "The development of children's motivation in school contexts", in P. D. Pearson. and A. Iran-Nejad (eds.), *Review of Research in Education* 23, American Educational Research Association, Washington D.C., pp. 73-118.
- Wilson, M.** (1994), "Comparing Attitude Across Different Cultures: Two Quantitative Approaches to Construct Validity", in M. Wilson (ed.), *Objective Measurement II: Theory into Practice*, Ablex, Norwood, pp. 271-292.
- Wolter, K.M.** (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.
- Wu, M.L., R.J. Adams and M.R. Wilson** (1997), *ConQuest: Multi-Aspect Test Software* [computer program], Australian Council for Education Research, Camberwell.
- Zimmerman, B.J. and D.H. Schunk** (eds.) (1989), *Self-Regulated Learning and Academic Achievement. Theory, Research and Practice*, Springer, New York.

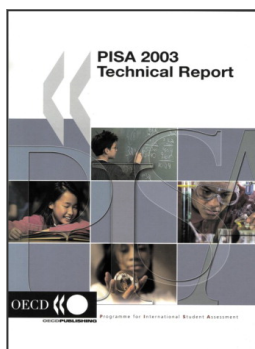


Table of Contents

Foreword	3
Chapter 1. The Programme for International Student Assessment: An overview	7
Reader's Guide	13
Chapter 2. Test design and test development	15
Chapter 3. The development of the PISA context questionnaires	33
Chapter 4. Sample design	45
Chapter 5. Translation and cultural appropriateness of the test and survey material	67
Chapter 6. Field operations	81
Chapter 7. Monitoring the quality of PISA	101
Chapter 8. Survey weighting and the calculation of sampling variance	107
Chapter 9. Scaling PISA cognitive data	119
Chapter 10. Coding reliability studies	135
Chapter 11. Data cleaning procedures	157
Chapter 12. Sampling outcomes	165
Chapter 13. Scaling outcomes	185
Chapter 14. Outcomes of coder reliability studies	217
Chapter 15. Data adjudication	235
Chapter 16. Proficiency scale construction	249
Chapter 17. Scaling procedures and construct validation of context questionnaire data	271
Chapter 18. International database	321
References	329



Appendix 1.	Sampling forms	335
Appendix 2.	PISA consortium and consultants	349
Appendix 3.	Country means and ranks by booklet.....	353
Appendix 4.	Item submission guidelines for mathematics – PISA 2003.....	359
Appendix 5.	Item review guidelines	379
Appendix 6.	ISCED adaptations for partner countries	383
Appendix 7.	Fictitious example of study programme table (SPT).....	389
Appendix 8.	Fictitious example of questionnaire adaptation spreadsheet (QAS).....	391
Appendix 9.	Summary of quality monitoring outcomes	393
Appendix 10.	Contrast coding for PISA 2003 conditioning variables	401
Appendix 11.	Scale reliabilities by country	409
Appendix 12.	Details of the mathematics items used in PISA 2003	411
Appendix 13.	Details of the reading items used in PISA 2003.....	415
Appendix 14.	Details of the science items used in PISA 2003	417
Appendix 15.	Details of the problem-solving items used in PISA 2003.....	419
Appendix 16.	Levels of parental education converted into years of schooling.....	421
Appendix 17.	Student listing form	423



From:
PISA 2003 Technical Report

Access the complete publication at:
<https://doi.org/10.1787/9789264010543-en>

Please cite this chapter as:

OECD (2006), "Test Design and Test Development", in *PISA 2003 Technical Report*, OECD Publishing, Paris.

DOI: <https://doi.org/10.1787/9789264010543-3-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.

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to rights@oecd.org. Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at info@copyright.com or the Centre français d'exploitation du droit de copie (CFC) at contact@cfcopies.com.