Programme for **I**nternational **S**tudent **A**ssessment

PISA 2003 Technical Report

OECD
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Foreword

The OECD's Programme for International Student Assessment (PISA) surveys, which take place every three years, have been designed to collect information about 15-year-old students in participating countries. PISA examines how well students are prepared to meet the challenges of the future, rather than how well they master particular curricula. The data collected during each PISA cycle are an extremely valuable source of information for researchers, policy makers, educators, parents and students. It is now recognised that the future economic and social well-being of countries is closely linked to the knowledge and skills of their populations. The internationally comparable information provided by PISA allows countries to assess how well their 15-year-old students are prepared for life in a larger context and to compare their relative strengths and weaknesses.

PISA is methodologically highly complex, requiring intensive collaboration among many stakeholders. The successful implementation of PISA depends on the use and sometimes further development, of state of the art methodologies and technologies. The *PISA 2003 Technical Report* describes those methodologies, along with other features that have enabled PISA to provide high quality data to support policy formation and review. The descriptions are provided at a level that will enable review and, potentially, replication of the implemented procedures and technical solutions to problems.

This report contains a description of the theoretical underpinning of the complex techniques used to create the PISA 2003 database, which contains information on over a quarter of a million students from 41 countries. The database includes not only information on student performance in the four main areas of assessment – reading, mathematics, science and problem solving – but also their responses to the student questionnaire that they complete as part of the assessment. Data from the school principals of participating schools are also included. The PISA 2003 database was used to generate information and to act as a base for analysis for the production of the PISA 2003 initial reports, *Learning for Tomorrow's World – First Results from PISA 2003* (OECD, 2004a) and *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003* (OECD, 2004b).

The information in this report complements the *PISA 2003 Data Analysis Manual: SAS® Users* (OECD, 2005a) and the *PISA 2003 Data Analysis Manual: SPSS® Users* (OECD, 2005b), which give detailed accounts of how to carry out the analyses of the information in the database.

PISA is a collaborative effort by the participating countries, and guided by their governments on the basis of shared policy-driven interests. Representatives of each country form the PISA Governing Board which decides on the assessment and reporting of results in PISA.

The OECD recognises the creative work of Raymond Adams of the Australian Council for Educational Research (ACER), who is project director of the PISA consortium and who acted as editor for this report, and his team, Alla Berezner, Eveline Gebhardt, Aletta Grisay, Marten Koomen, Sheila Krawchuk, Christian Monseur, Martin Murphy, Keith Rust, Wolfram Schulz, Ross Turner and Norman Verhelst. A full list of the contributors to the PISA project is included in Appendix 2 of this report. The editorial work at the OECD Secretariat was carried out by John Cresswell, Miyako Ikeda, Sophie Vayssettes, Claire Shewbridge and Kate Lancaster.

Table of Contents

1

The Programme for International Student Assessment: An Overview

The OECD's Programme for International Student Assessment (PISA) is a collaborative effort among OECD member countries to measure how well 15-year-old young adults approaching the end of compulsory schooling are prepared to meet the challenges of today's knowledge societies. The assessment is forward-looking: rather than focusing on the extent to which these students have mastered a specific school curriculum, it looks at their ability to use their knowledge and skills to meet real-life challenges. This orientation reflects a change in curricular goals and objectives, which are increasingly concerned with what students can do with what they learn at school.

The first PISA survey was conducted in 2000 in 32 countries (including 28 OECD member countries) using written tasks answered in schools under independently supervised test conditions. Another 11 countries completed the same assessment in 2002. PISA 2000 surveyed reading, mathematical and scientific literacy, with a primary focus on reading.

The second PISA survey, which covered reading, mathematical and scientific literacy, and problem solving, with a primary focus on mathematical literacy, was conducted in 2003 in 41 countries. This report is concerned with the technical aspects of this second PISA survey, which is usually referred to as PISA 2003.

In addition to the assessments, PISA 2003 included Student and School Questionnaires to collect data that could be used in constructing indicators pointing to social, cultural, economic and educational factors that are associated with student performance. Using the data taken from these two questionnaires, analyses linking context information with student achievement could address:

- Differences between countries in the relationships between student-level factors (such as gender and social background) and achievement;
- Differences in the relationships between school-level factors and achievement across countries;
- Differences in the proportion of variation in achievement between (rather than within) schools, and differences in this value across countries;
- Differences between countries in the extent to which schools moderate or increase the effects of individual-level student factors and student achievement;
- Differences in education systems and national contexts that are related to differences in student achievement across countries; and
- Through links to PISA 2000, changes in any or all of these relationships over time.

Through the collection of such information at the student and school level on a cross-nationally comparable basis, PISA adds significantly to the knowledge base that was previously available from national official statistics, such as aggregate national statistics on the educational programs completed and the qualifications obtained by individuals.

The ambitious goals of PISA come at a cost: PISA is both resource intensive and methodologically complex, requiring intensive collaboration among many stakeholders. The successful implementation of PISA depends on the use, and sometimes the further development, of state-of-the-art methodologies.

This report describes some of those methodologies, along with other features that have enabled PISA to provide high quality data to support policy formation and review. Figure 1.1 provides an overview of the central design elements of PISA 2003. The remainder of this report describes these design elements, and the associated procedures, in more detail.

Figure 1.1 ■ **Core features of PISA 2003**

Sample size

- More than a quarter of a million students, representing almost 30 million 15-year-olds enrolled in the schools of the 41 participating countries, were assessed in 2003.

Content

- PISA 2003 covered four domains: reading literacy, mathematical literacy, scientific literacy and problem solving.
- PISA 2003 looked at young people's ability to use their knowledge and skills in order to meet real-life challenges rather than how well they had mastered a specific school curriculum. The emphasis was placed on the mastery of processes, the understanding of concepts, and the ability to function in various situations within each domain.

Methods

- PISA 2003 used paper-and-pencil assessments, lasting two hours for each student.
- PISA 2003 used both multiple-choice items and questions requiring students to construct their own answers. Items were typically organised in units based on a stimulus presenting a real-life situation.
- A total of six and a half hours of assessment items was created, with different students taking different combinations of the assessment items.
- Students answered a background questionnaire that took about 30 minutes to complete and, as part of international options, completed questionnaires on their educational careers as well as familiarity with computers.
- School principals completed a questionnaire about their school.

Outcomes

- A profile of knowledge and skills among 15-year-olds.
- Contextual indicators relating results to student and school characteristics.
- A knowledge base for policy analysis and research.
- Trend indicators showing how results change over time.

MANAGING AND IMPLEMENTING PISA

The design and implementation of PISA 2003 was the responsibility of an international consortium led by the Australian Council for Educational Research (ACER). The other partners in this consortium were the National Institute for Educational Measurement (CITO) in the Netherlands, Westat and the Educational Testing Service (ETS) in the United States, and the National Institute for Educational Research (NIER) in Japan. Appendix 2 lists the many consortium staff and consultants who have made important contributions to the development and implementation of the project.

The consortium implements PISA within a framework established by the PISA Governing Board (PGB), which includes representation from all countries at senior policy levels. The PGB established policy priorities and standards for developing indicators, for establishing assessment instruments, and for reporting results. Experts from participating countries served on working groups which linked the programme policy objectives with the best internationally available technical expertise in the four assessment areas. These expert groups were referred to as subject matter expert groups (SMEGs) (see Appendix 2 for members). By participating in these expert groups and regularly reviewing outcomes of the groups' meetings, countries ensured that the instruments were internationally valid and took into account the cultural and educational contexts of the different OECD member countries; the assessment materials had strong measurement potential; and that the instruments emphasised authenticity and educational validity.

Participating countries implemented PISA nationally through National Project Managers (NPMs), who respected common technical and administrative procedures. These managers played a vital role in developing and validating the international assessment instruments and ensured that PISA implementation was of high quality. The NPMs also contributed to the verification and evaluation of the survey results, analyses and reports.

The OECD Secretariat had overall responsibility for managing the programme. It monitored its implementation on a day-to-day basis, served as the secretariat for the PGB, fostered consensus building between the countries involved, and served as the interlocutor between the PGB and the international consortium.

THIS REPORT

This Technical Report does not report the results of PISA. The first results from PISA 2003 were published in December 2004 in *Learning for Tomorrow's World – First Results from PISA 2003* (OECD, 2004a) and *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003* (OECD, 2004b).

This Technical Report is designed to describe the technical aspects of the project at a sufficient level of detail to enable review and, potentially, replication of the implemented procedures and technical solutions to problems. The report covers five main areas:

- *Instrument Design:* Covers the design and development of both the achievement tests and questionnaires.
- *Operations:* Covers the operational procedures for the sampling and population definitions, test administration procedures, quality monitoring and assurance procedures for test administration and national centre operations, and instrument translation.
- *Data Processing:* Covers the methods used in data cleaning and preparation, including the methods for weighting and variance estimation, scaling methods, methods for examining inter-rater variation and the data cleaning steps.

- *Quality Indicators and Outcomes:* Covers the results of the scaling and weighting, reports response rates and related sampling outcomes and gives the outcomes of the inter-rater reliability studies. The last chapter in this section summarises the outcomes of the PISA 2003 data adjudication – that is, the overall analysis of data quality for each country.
- *Scale Construction and Data Products:* Describes the construction of the PISA 2003 described levels of proficiency and the construction and validation of questionnaire-related indices. The final chapter briefly describes the contents of the PISA 2003 database.
- *Appendices:* Detailed appendices of results pertaining to the chapters of the report are provided.

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

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.

© OECD 2005

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 panelling, 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 forum[1] 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)**

	Competency cluster			
Item format	**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)**

	Competency cluster			
Content category	**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)**

	Item format			
Content category	**Multiple-choice response**	**Closed-constructed response**	**Open-constructed response**	**Total**
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)**

	Item format			
Problem-solving type	**Closed-constructed response**	**Multiple-choice response**	**Open-constructed response**	**Total**
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)**

	Item format					
Science process	**Closed-constructed response**	**Complex multiple-choice response**	**Multiple-choice response**	**Open-constructed response**	**Short response**	**Total**
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 category[2]. 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)**

	Competency cluster			
Item format	**Reproduction**	**Connections**	**Reflection**	**Total**
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**

© OECD 2005

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

	Competency cluster			
Content category	**Reproduction**	**Connections**	**Reflection**	**Total**
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)**

	Item format			
Content category	**Multiple-choice response**	**Closed-constructed response**	**Open-constructed response**	**Total**
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)**

	Item format				
Reading process	**Closed-constructed response**	**Multiple-choice response**	**Open-constructed response**	**Short response**	**Total**
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)**

	Item format				
Text structure type	**Closed-constructed response**	**Multiple-choice response**	**Open-constructed response**	**Short response**	**Total**
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)**

	Item format			
Problem-solving type	**Closed-constructed response**	**Multiple-choice response**	**Open-constructed response**	**Total**
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)**

	Item format				
Science process	**Complex-multiple choice**	**Multiple-choice response**	**Open-constructed response**	**Short response**	**Total**
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.

© OECD 2005

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.

3

The Development of the PISA Context Questionnaires

OVERVIEW

In addition to the assessment of the achievement of 15-year-old students in reading, science, mathematics, and problem-solving skills, PISA 2003 also included the collection of information on the characteristics of students and their schools. This was done with the aim of identifying social, cultural, economic and educational factors that are associated with student performance. For this purpose student and school questionnaires were completed by the students and the principals of the sampled schools. In addition to a core student questionnaire, two internationally optional student questionnaires, the Information Communication Technology Familiarity and Educational Career questionnaires, were offered to participating countries. Using the data from these context questionnaires, analyses linking context information with student outcomes allows PISA to examine:

- Differences between countries in the relationships between student-level factors (such as gender and social background) and outcomes;
- Differences in the relationships between school-level factors and outcomes across countries;
- The proportion of variation in outcomes between (rather than within) schools, and differences in this value across countries;
- Differences between countries in the extent to which schools moderate or increase the effects of individual-level student factors and student outcomes;
- Characteristics of education systems and national contexts that are related to differences in student outcomes across countries; and
- Changes in any or all of these relationships over time.

The PGB requested that PISA 2003 portray important aspects of learning and instruction in mathematics, including the impact of learning and teaching strategies on achievement, as well as the impact of school organisation and structures in promoting active student engagement with learning. Furthermore, the PGB requested that PISA 2003 address issues related to mathematics efficacy and students' engagement, motivation and confidence with mathematics, mathematics and gender, and students' planned educational pathways. Finally, the quality of the school's human and material resources, issues of public/private control, management and funding, school level information on the instructional context and institutional structures were also considered important issues in PISA 2003.

To accomplish these goals, the following steps were taken:

- First, an organising framework was established that allowed the mapping of these policy issues against the design and instrumentation of PISA. The objective was to facilitate choosing research areas that combine policy relevance effectively with the strengths of the PISA[1] design.
- After a conceptual structure from which relevant research focus areas or themes could be established was identified, a set of criteria was developed for defining and operationalising the PGB's policy priorities within this conceptual structure.
- Third, proposals for potential thematic reports for PISA 2003 were outlined, with each proposal presenting a brief review of relevant literature, the specific policy questions the report could address, and how this would be operationalised in the PISA 2003 context questionnaires.

THE CONCEPTUAL FRAMEWORK

To facilitate a systematic approach to the organisation and prioritisation of research focus areas, the framework for the OECD education indicators (INES) was applied. The INES framework organises policy issues that might be considered in PISA by using two dimensions:

- The level of the education system to which the resulting indicators relate; and
- Whether they relate to outcomes or outputs, policy-amenable determinants of these outcomes or outputs or constraints at the respective level of the education system.

The INES framework considered four levels that related both to the entities from which data might be collected and to the recognition that national education systems are multi-levelled. The four levels are:

- The education system as a whole;
- The educational institutions and providers of educational services;
- The instructional setting and the learning environment within the institutions; and
- The individual participants in learning activities.

A differentiation between levels is not only important with regard to the collection of information, but also because many features of the education system play out quite differently at different levels of the system. For example, at the level of the students within a classroom, the relationship between student achievement and class size may be negative, if students in small classes benefit from improved contact with teachers. At the class or school level, however, students are often intentionally grouped such that weaker or disadvantaged students are placed in smaller classes so that they receive more individual attention. At the school level, therefore, the observed relationship between class size and student achievement is often positive (suggesting that students in larger classes perform better than students in smaller classes). At higher aggregated levels of education systems, the relationship between student achievement and class size is further confounded, e.g. by the socio-economic intake of schools or by factors relating to the learning culture in different countries. Past analyses, which have relied on macro-level data alone, have therefore sometimes led to misleading conclusions.

The second dimension in the organising framework further groups the indicators at each of the above levels (*i.e.* system, institutional, classroom or individual) under the following subheadings:

- *Output and outcomes of education and learning:* Indicators on observed outputs of education systems, as well as indicators related to the impact of knowledge and skills for individuals, societies and economies.
- *Policy levers and contexts:* Activities seeking information on the policy levers or circumstances that shape the outputs and outcomes at each level.
- *Antecedents and constraints:* Policy levers and contexts typically have antecedents, that is, factors that define or constrain policy. It should be noted that the antecedents or constraints are usually specific for a given level of the education system, and that antecedents at a lower level of the system may well be policy levers at a higher level (*e.g.* for teachers and students in a school, teacher qualifications are a given constraint while, at the level of the education system, professional development of teachers is a key policy lever).

This basic conceptualisation has been adapted from the conceptual framework for the Second IEA Study of Mathematics (Travers and Westbury, 1989; Travers *et al.*, 1989) and also provided a conceptual basis for the planning of context questionnaires in PISA 2000 (see Harvey-Beavis, 2002). Figure 3.1 shows the two-dimensional matrix of the four levels and the three aspects. Each cell also contains a description of data that were eventually collected in PISA 2003.

While this mapping is useful for describing the coverage of the PISA questionnaires, it is also important to supplement it with the recognition of the dynamic elements of the education system. All of the cells in the framework are linked, both directly and indirectly, and a range of important indicators that deal with the relations between the cells are central to the outcomes of PISA 2003. For example, analysing the impact of socio-economic background on student performance is directly concerned with the relationship between cells 9 and 1, and at the same time its further exploration is concerned with how data relating to cells 5 to 8 might influence this relationship.

Because PISA 2003 did not survey teachers, nor had intact classrooms as units of sampling, there are limits on the availability and relevance of data on some classroom contexts and antecedents, such as teacher characteristics and qualifications, and on classroom processes such as pedagogical practices and curriculum content (cells 2, 6 and 10). Any information on these aspects could only be collected either from students or at the school level. Therefore, the data collected on classroom processes (cell 6) refer to the classroom practices but are collected from students learning in different instructional settings across the school and can only be analysed at the student or school level.

Similarly, at the school level (cells 7 and 11), PISA focused on questions that were related to relatively broad and stable features such as school type, school structure, school resources, school climate and school management, most of which are known to have some impact on student's achievement, according to the school effectiveness literature (see Teddlie and Reynolds, 2000).

PISA 2003 did not include any activities that directly focused on collecting data at the national level as included in cells 8 and 12. A range of such data is however available from the OECD education indicators programme and can be included in the analysis of the database.

RESEARCH THEMES IN PISA 2003

To capitalise on the PISA design and to maximise the contributions PISA could make to the needs of policy makers and educators it was important to choose wisely from the wide range of possible policy-relevant research themes.

The definition and operationalisation of policy-relevant research areas for potential thematic reports was guided by the following requirements which were developed by OECD INES Network A:

- First, they had to be of enduring policy relevance and interest. A research focus area needed to have policy relevance, to capture policy makers' attention, to address their needs for data about the performance of their educational systems, to be timely, and to focus on factors that improve or explain the outcomes of education. Further, a theme had to be of interest to the public, since it is this public to which educators and policy makers are accountable.
- Second, the themes were to provide an internationally comparative perspective, and add significant value beyond that which can be accomplished through national evaluation and analysis. This implies that

Figure 3.1 ■ **Mapping of PISA 2003 data to conceptual grid**

	Column 1 Outputs and outcomes of education and learning	Column 2 Policy levers and contexts	Column 3 Antecedents and constraints
Individual participants in education and learning	*Cell 1:* ***Individual outcomes*** *Student test data collected in 2003:* - Reading, mathematics and science literacy - Problem-solving skills *Student Questionnaire data collected in 2003:* - Self-related cognitions in mathematics (self-efficacy, self-concept) - Motivational factors: interest in and enjoyment of mathematics - Educational expectations	*Cell 5:* ***Policy levers and contexts relating to individuals*** *Student Questionnaire data collected in 2003:* - Students' perception of school (student/teacher relations, sense of belonging, attitudes toward school) - Learning strategies and preferences - Instrumental motivation to learn mathematics - Emotional factors (mathematics anxiety) - Instructional time - Study time in mathematics and other subjects (homework, extension/remedial, tutoring, out-of-school classes, other activities)	*Cell 9:* ***Antecedents and constraints at the level of individuals*** *Student Questionnaire data collected in 2003:* - Home possessions - Parental education - Parental occupation - Family structure - Country of birth - Language spoken at home - Age and gender - Grade and study programme - Prior education (pre-schooling, entry age, retention)
Instructional settings	*Cell 2:* ***Outputs and outcomes at the level of classrooms/instructional settings***	*Cell 6:* ***Policy levers and contexts at the level of classrooms/instructional settings*** *Student Questionnaire data collected in 2003:* - Disciplinary climate in mathematics lessons (student perceptions) - Teacher support in mathematics lessons (student perceptions) - Use of textbooks in mathematics lessons (student perceptions) - Classroom size (student perceptions)	*Cell 10:* ***Antecedents and constraints at the level of classrooms/instructional settings***
Education service providers	*Cell 3:* ***Outputs and outcomes at the level of institutions*** *Data available in 2003:* - Aggregates of cell 1 (literacy scores, motivation and self-related cognitions)	*Cell 7:* ***Policy levers and contexts at the level of institutions*** *School Questionnaire data collected in 2003:* - School resources (quality of human, educational and material resources, teacher and computer availability) - Admittance and grouping policies - Curricular practices (mathematics activities, student assessment, retention, instructional time, monitoring of teachers) - School climate (student/teacher behaviour, student/teacher morale)	*Cell 11:* ***Antecedents and constraints at the level of institutions*** *School Questionnaire data collected in 2003:* - The type of school, its source of funding, its location and size (students and grade levels) - Language background of students - Responsibilities for decision making *Student Questionnaire data aggregated to the school level:* - Socio-economic background of students (intake)
The education system as a whole	*Cell 4:* ***Outcomes at the level of the education system*** *Data available in 2003:* - System-level aggregates of cell 1 - Equity related outcomes	*Cell 8:* ***Policy levers and contexts at the national level*** *Data available in 2003:* - System-level aggregates from cell 7 - OECD data	*Cell 12:* ***Macro-economic and demographic context*** *Data available in 2003:* - System-level aggregates from cell 7 - OECD data

themes needed to be both relevant (that is, of importance) and valid (that is, of similar meaning) across countries.

- Third, there had to be consistency in approach and themes with PISA 2000.
- Fourth, the implementation of a research focus area had to be technically feasible and appropriate within the context of the PISA design. That is, the collection of data about a subject needed to be technically feasible in terms of methodological rigour and the time and costs (including opportunity costs) associated with data collection.

The following proposals for thematic reports were elaborated for PISA 2003 in accordance with the priorities established by the PGB and the criteria outlined above:

- *School characteristics, organisation and structure:* PISA 2003 provided an opportunity to explore some key variables that might cause variance between schools. These variables where grouped into variables related primarily to the structure of schooling (ability grouping, segregation of schools, management and financing, school resources, size and location) and those related to the instructional context within schools (learning time, student support policies, school and classroom climate).
- *Teaching and learning strategies:* Theoretical and empirical research on teacher instructional practices, student learning strategies and the impact of such variables on student achievement is extensive. Given the design of PISA, which does not include a classroom level of analysis, priority was given to dimensions that might reasonably be considered as being pervasive characteristics of either the instructional context or of students' learning strategies.
- *Student engagement with mathematics:* Students' engagement with learning is crucial for the acquisition of proficiency, and is also an important outcome of education. Students' engagement refers to both students active involvement in learning, and to students' beliefs about their own ability to succeed in a subject, motivation to learn a subject and emotional relationship with a subject, as well as their choice of learning strategies for a subject. This theme covers the following aspects of engagement with mathematics: Self-related cognitions, motivational preferences, emotional factors and behaviour-related variables.
- *Mathematics and gender:* Gender differences in achievement are ongoing equity related concerns in OECD countries, and as such, are central to PISA. Given the focus of PISA 2003 on mathematics this theme addresses gender differences in mathematics literacy, differences in mathematics-related attitudes and self-cognitions, and career expectations.
- *Students' educational career:* One of the challenges faced by educational systems is to ensure that, although learning takes place in collective settings (schools, classrooms), the individual needs of learners are served in an efficient way. This theme addresses issues related to how educational systems shape educational careers of students and to what extent they influence students' career expectations.
- *Use of and access to technology:* This theme is linked to the ICT familiarity international option and addresses issues such as the availability of ICT at schools, the students' familiarity (use, self-confidence and attitudes) and the role of ICT in the instructional context.
- *Family background and student performance:* Educational outcomes are influenced by family background in many different and complex ways. In particular, the socio-economic status of families has been consistently found to be an important variable in explaining variance in student achievement. This theme addresses the impact of socio-economic background, ethnicity (language and immigrant background) and family structure on student performance.

In the elaboration of these research areas, variables or constructs were identified which needed to be included in the context questionnaires. Table 3.1 details the major constructs and variables identified as important within each of these research themes. Some of these constructs or variables form the core of the questionnaire material, which remains unchanged across PISA cycles. The core component comprises questions about basic school or student characteristics and the students' socio-economic background.

Table 3.1 ■ **Themes and constructs in PISA 2003**

Research theme	Constructs (or variables)
School characteristics, organisation and structure	School size, location and funding
	Language background and school policies
	Quality of school resources (staff, educational material)
	Admittance policies
	Ability grouping
	Assessment practices
	Activities to promote engagement with mathematics
	Teacher morale
	Student morale
	Teacher behaviour
	Student behaviour
	Mathematics teacher agreement or dissent
	School autonomy in decision making
	Influence on decision making by school-related groups
Teaching and learning strategies	Learning strategies (memorisation, control, elaboration)
	Learning style preferences (co-operative, competitive)
	Classroom climate (disciplinary climate, teacher support)
Student engagement with mathematics	Mathematics self-efficacy
	Mathematics self-concept
	Mathematics anxiety
	Interest in and enjoyment of mathematics
	Instrumental motivation to learn mathematics
	Study time in mathematics
Mathematics and gender	Gender
Students' educational career	Pre-school attendance
	School entry age
	Grade repetition
	Expected educational level
	Retention rate at school
Use and access to technology	Use of and experience with computers
	Types of ICT use (Internet/entertainment, programme use)
	Self-confidence in ICT (routine, Internet, high-level)
	Attitudes toward computers
	Source of ICT knowledge
	Availability of computers at school
Family background and student performance	Immigrant background
	Language use
	Home possessions (cultural, educational)
	Parental occupation
	Parental education
	Family structure

THE DEVELOPMENT OF THE CONTEXT QUESTIONNAIRES

The development of questionnaire material was guided by the PGB priorities and their elaboration in the conceptual framework. Some of the questionnaire items used in PISA 2000 were retained: some because they were considered as a core part of the context questionnaires and will be included in each cycle, others because they were important for the analyses proposed as part of the research focus areas.

However, many of the constructs or variables were new, and were developed during the two years prior to the assessment. The new questionnaire material was developed in co-operation with international experts, the OECD and national centres.

After an initial phase of piloting questionnaire material in a few participating countries, to look at qualitative as well as some quantitative aspects of item responses, a final draft version of the material was presented to national centres. After extensive consultations with national centres, international centres and the OECD, two different student questionnaire versions and a school questionnaire were administered in a field trial in all participating countries. Each questionnaire version included, in addition to a set of common items, different questions plus common questions trial-tested with a different item format.

The questionnaires were trialled together with the achievement test in 2002 in all participating countries. The data analysis of the field trial data included the following steps:

- An examination of non-response and response patterns for the questionnaire items;
- A comparison of different item formats between the two versions of the questionnaire;
- Exploratory and confirmatory factor analysis to review the dimensional structure of questionnaire items and to facilitate the selection of constructs and items;
- An analysis of cross-country validity of both dimensional item structure and item fit (student-level data only); and
- A review of scaling properties for scaled items, using classical item statistics and IRT models.

Analyses of the field trial data were carried out in the second half of 2002 and a proposal of final questionnaire material for the main study was developed based on these results. The final selection of questionnaire material was made after an extensive review and consultations with national centres, international experts and the OECD. The selection process was principally based on the following criteria:

- Scaling properties of items used to measure constructs;
- Predictive validity of constructs;
- Cross-cultural appropriateness of the material; and
- Priority judgements about constructs and items in accordance with questionnaire framework and the policy issues set by the PGB.

THE COVERAGE OF THE QUESTIONNAIRE MATERIAL

Student Questionnaire

In the main study the student questionnaire was administered after the assessment and it took students about 35 minutes to complete. The questionnaire covered the following aspects:

- *Student characteristics:* Grade, study programme, age and gender.
- *Family background:* Family structure, employment status of parents, occupation of parents, education of parents, home possessions, number of books at home, country of birth for student and parents, language spoken at home.
- *Educational background of student:* Pre-schooling, primary school starting age, grade repetition, expected education, attitudes toward school in general.
- *Student reports related to the school:* Reasons for selecting school, student-teacher relations, sense of belonging to school, late arrivals at school, study time for all subjects (homework, school extension courses, out-of-school classes, tutoring, other study).
- *Students' learning of mathematics:* Interest in and enjoyment of mathematics, instrumental motivation to learn mathematics, mathematics self-efficacy, mathematics self-concept, mathematics anxiety, study time for mathematics (homework, school extension courses, out-of-school classes, tutoring, other study) and learning strategies in mathematics (memorisation, elaboration and control strategies).
- *Students' lessons in mathematics:* Instructional time (mathematics, overall), preference for learning situations (competitive, co-operative), classroom climate (teacher support, disciplinary climate).

School Questionnaire

The main study school questionnaire was administered to the school principal and took about 20 minutes to complete. It covered a variety of school-related aspects:

- *School characteristics:* Community size, enrolment, ownership, funding and number of grade levels at school.
- *The school's resources:* Instructional time, quality of resources (staffing, educational material, infrastructure) and computers available at school.
- *The student body:* Student admittance criteria, student morale, language background of students, student behaviour and grade repetition.
- *Teachers in the school:* Staffing, monitoring of teachers, principal perceptions of consistent and shared goals among mathematics staff, teacher morale and teacher behaviour.
- *Pedagogical practices of the school:* Activities to promote student learning of mathematics, ability grouping, student assessments, use of assessments and foreign language courses.
- *Administrative structures within the school:* Responsibilities for decision making at school and bodies influencing decision making at school.

International options

As in PISA 2000, additional questionnaire material was developed and offered as international options to participating countries. In PISA 2003, two international options were available: the ICT Familiarity questionnaire and Educational Career Questionnaire.

Educational Career Questionnaire

The inclusion of an optional Educational Career questionnaire was due to the fact that not all of the participating countries expressed interest in this particular research area. National centres were allowed to select any of the items included in this questionnaire for inclusion without having to administer all of the questions. The completion of this questionnaire took about two minutes and covered the following aspects:

- *Past educational career:* Missing of school at primary and lower secondary level, change of school at primary and lower secondary level, change of study programme.
- *Present educational settings:* Difficulty level of current mathematics course or lessons, teacher marks in mathematics.
- *Expected occupation.*

Information Communication Technology Questionnaire

The Information Communication Technology (ICT) questionnaire consisted of questions regarding the students' use of, familiarity with, and attitudes towards ICT. ICT was defined as the use of any equipment or software for processing or transmitting digital information that performs diverse general functions, whose options can be specified or programmed by its user. The questionnaire was administered to students after the international student questionnaire (sometimes combined within the same booklet) and it took about five minutes to complete. It covered the following aspects:

- *Use of ICT:* Availability of computers, students' experience with computers and location of use, frequency of ICT for different purposes;
- *Affective responses to ICT:* Self-confidence with ICT (routine, Internet and high-level programming tasks), attitudes towards computers; and
- *Learning of ICT:* Sources of students' ICT and Internet knowledge.

National questionnaire material

National centres could decide to add national items to the international student or school questionnaire. Insertion of national items into the student questionnaire had to be agreed upon with the international study centre during the review of adaptations, due to context relatedness. Adding more than five national items was considered as a national option. National student questionnaire options, which took less than ten minutes to be completed, could be administered after the international student questionnaire and international options. If the length of the national options exceeded ten minutes, national centres were requested to administer their national questionnaire material in follow-up sessions.

THE IMPLEMENTATION OF THE CONTEXT QUESTIONNAIRES

In order to ensure that all questions were understood by 15-year-old students and school principals in all participating countries, it was necessary to adapt parts of the questionnaire material from the international source version to the national context. Such adaptations had to be carefully monitored so that the comparability of the collected data was not jeopardised. This is particularly important with questions that relate to the educational system such as educational levels, study programmes or certain school characteristics which differ in terminology across countries.

To achieve maximum comparability, a process was implemented during which each adaptation was reviewed and discussed by the international study centre and national study centres. To facilitate this process, national centres were asked to complete a questionnaire adaptation spreadsheet (QAS, see Appendix 8), where adaptations to the questionnaire material were documented.

Each adaptation had to be reviewed and agreed upon before the questionnaire material could be submitted for translation verification and the final optical check (see Chapter 5). The QAS also contained information about additional national questionnaire material and any deviation from the international questionnaire format, as well as the corresponding variable names in the national database, which was submitted after data collection.

Prior to the review of questionnaire adaptations, national centres had been asked to complete Study Programme Tables (SPT, see Appendix 5) in order to document the range of different study programmes that are available for 15-year-old students across participating countries. This information was used as a codebook to collect these data from school records and also assisted the review of questionnaire adaptations.

Information on parental occupation and the students' expected occupation was collected through open-ended questions. The responses were then coded according to the International Standard Classification of Occupations (ISCO) (ILO, 1990). Once occupations had been coded into ISCO, the codes were re-coded into the International Socio-Economic Index of Occupational Status (ISEI) (Ganzeboom *et al.,* 1992), which provides a measure of the socio-economic status of occupations comparable across the countries participating in PISA.

The International Standard Classification of Education (ISCED) was used as a typology to classify educational qualifications and study programmes. The ISCED classification was used to obtain comparable data across countries. Whereas this information was readily available for OECD member countries,[2] for partner countries extensive reviews of their educational systems in co-operation with national centres were necessary to map educational levels to the ISCED framework (see Appendix 6).

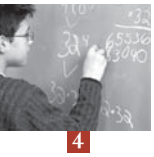
Notes

1 The questionnaire framework was not published by the OECD but is available as a project working document TAG(0303)4.doc

2 Partner countries are non-OECD member countries that participate in PISA.

4

Sample Design



TARGET POPULATION AND OVERVIEW OF THE SAMPLING DESIGN

The desired base PISA target population in each country consisted of 15-year-old students attending educational institutions located within the country, in grades 7 and higher. This meant that countries were to include 15-year-olds enrolled full-time in educational institutions, 15-year-olds enrolled in educational institutions who attended on only a part-time basis, students in vocational training types of programmes, or any other related type of educational programmes, and students attending foreign schools within the country (as well as students from other countries attending any of the programmes in the first three categories). It was recognised that no testing of persons schooled in the home, workplace or out of the country would occur and therefore these students were not included in the international target population.

The operational definition of an age population directly depends on the testing dates. The international requirement was that each country had to choose a 42-day period, referred to as the testing window, between 1 March 2003 and 31 August 2003, during which they would administer the assessment.

Further, testing was not permitted during the first three months of the school year because of a concern that student performance levels even after controlling for age may be lower at the beginning of the academic year than at the end of the previous academic year.

The 15-year-old international target population was slightly adapted to better fit the age structure of most of the northern hemisphere countries. As the majority of the testing was planned to occur in April, the international target population was consequently defined as all students aged from 15 years and 3 (completed) months to 16 years and 2 (completed) months at the beginning of the assessment period. This meant that in all countries testing in April 2003, the national target population could have been defined as all students born in 1987 who were attending a school or other educational institution.

Further, a variation of up to one month in this age definition was permitted. For instance, a country testing in March or in May was still allowed to define the national target population as all students born in 1987. If the testing was to take place at another time, the birth date definition had to be adjusted and approved by the consortium.

The sampling design used for the PISA assessment was a two-stage stratified sample in most countries. The first-stage sampling units consisted of individual schools having 15-year-old students. In all but a few countries, schools were sampled systematically from a comprehensive national list of all eligible schools with probabilities that were proportional to a measure of size. This is referred to as probability proportional to size (PPS) sampling.

The measure of size was a function of the estimated number of eligible 15-year-old students enrolled. Prior to sampling, schools in the sampling frame were assigned to strata formed either explicitly or implicitly. The second-stage sampling units in countries using the two-stage design were students within sampled schools. Once schools were selected to be in the sample, a list of each sampled school's 15-year-old students was prepared. From each list that contained more than 35 students, 35 students were selected with equal probability, and for lists of fewer than 35, all students on the list were selected. It was possible for countries to sample a number of students within schools other than 35, provided that the number sampled within each school was at least as large as 20.

In two countries, a three-stage design was used. In such cases, geographical areas were sampled first (called first-stage units) using probability proportional to size sampling, and then schools (called second-stage units) were selected within sampled areas. Students were the third-stage sampling units in three-stage designs.

POPULATION COVERAGE, AND SCHOOL AND STUDENT PARTICIPATION RATE STANDARDS

To provide valid estimates of student achievement, the sample of students had to be selected using established and professionally recognised principles of scientific sampling, in a way that ensured representation of the full target population of 15-year-old students.

Furthermore, quality standards had to be maintained with respect to the coverage of the international target population, accuracy and precision, and the school and student response rates.

Coverage of the PISA international target population

In an international survey in education, the types of exclusion must be defined internationally and the exclusion rates have to be limited. Indeed, if a significant proportion of students were excluded, this would mean that survey results would not be deemed representative of the entire national school system. Thus, efforts were made to ensure that exclusions, if they were necessary, were minimised.

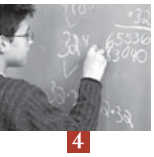
Exclusion can take place at the school level (the whole school is excluded) or at the within-school level. In PISA, there are several reasons why a school or a student can be excluded. Exclusions at school level might result from removing a small, remote geographical region due to inaccessibility or size, or from removing a language group, possibly due to political, organisational or operational reasons. Areas deemed by the PISA Governing Board (PGB) to be part of a country (for the purpose of PISA), but which were not included for sampling, were designated as non-covered areas, and documented as such – although this occurred infrequently. Care was taken in this regard because, when such situations did occur, the national desired target population differed from the international desired target population.

International within-school exclusion rules for students were specified as follows:

- Intellectually disabled students are students who are considered in the professional opinion of the school principal, or by other qualified staff members, to be intellectually disabled, or who have been tested psychologically as such. This category includes students who are emotionally or mentally unable to follow even the general instructions of the test. Students were not to be excluded solely because of poor academic performance or normal discipline problems.
- Functionally disabled students are students who are permanently physically disabled in such a way that they cannot perform in the PISA testing situation. Functionally disabled students who could perform were to be included in the testing.
- Students with limited proficiency in the language of the PISA test were excluded if they had received less than one year of instruction in the language(s) of the test.

A school attended only by students who would be excluded for intellectual, functional or linguistic reasons was considered as a school-level exclusion.

It was required that the overall exclusion rate within a country be kept below 5 per cent. Restrictions on the level of exclusions of various types were as follows:



- School-level exclusions for inaccessibility, feasibility or other reasons were required to cover fewer than 0.5 per cent of the total number of students in the international PISA target population. Schools on the school sampling frame that had only one or two eligible students were not allowed to be excluded from the frame. However, if, based on the frame, it was clear that the percentage of students in these schools would not cause a breach of the 0.5 per cent allowable limit, then such schools could be excluded in the field, if at that time, they still only had one or two PISA eligible students. This procedure was changed from PISA 2000 to increase coverage by guarding against any such schools possibly having three or more eligible students at the time of data collection.
- School-level exclusions for intellectually or functionally disabled students, or students with limited proficiency in the language of the PISA test, were required to cover fewer than two per cent of students.
- Within-school exclusions for intellectually disabled or functionally disabled students or students with limited language proficiency were required to cover fewer than 2.5 per cent of students. However, if the percentage was greater than 2.5 per cent, it was re-examined without including the students excluded because of limited language proficiency, since this is a largely unpredictable part of each country's eligible population.

Accuracy and precision

A minimum of 150 schools (or all schools if there were fewer than 150 schools in a participating jurisdiction) had to be selected in each country. Within each participating school, a sample of the PISA eligible students was selected with equal probability. The within-school sample size (sometimes referred to as the "target cluster size") was usually 35 students. In schools where there were fewer eligible students than the target cluster size, all students were sampled. In total, a minimum sample size of 4 500 assessed students was to be achieved. It was possible for countries to negotiate a different target cluster size, but if it was reduced then the sample size of schools was increased beyond 150, so as to ensure that at least 4 500 students in total would be assessed. The target cluster size had to be at least 20 so as to ensure adequate accuracy in estimating variance components within and between schools – an analytical objective of PISA.

National Project Managers (NPMs) were strongly encouraged to identify stratification variables to reduce the sampling variance.

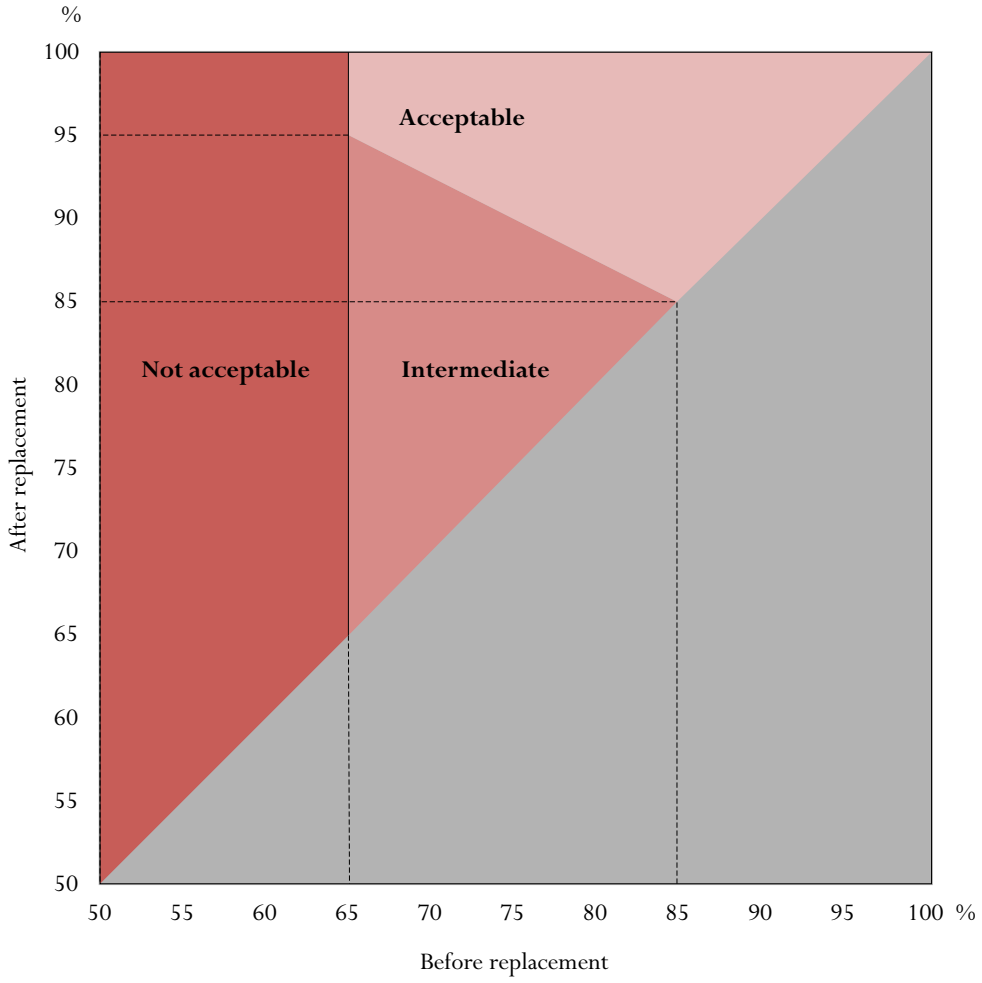
For countries that had participated in PISA 2000 and that had larger than anticipated sampling variances associated with their estimates, recommendations were made about sample design changes that would help to reduce the sampling variances for PISA 2003. These included modifications to stratification variables, and increases in the required sample size.

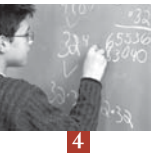
School response rates

A response rate of 85 per cent was required for initially selected schools. If the initial school response rate fell between 65 and 85 per cent, an acceptable school response rate could still be achieved through the use of replacement schools. Figure 4.1 provides a summary of the international requirements for school response rates. To compensate for a sampled school that did not participate, where possible two replacement schools were identified for each sampled school. Furthermore, a school with a student participation rate between 25 and 50 per cent was not considered as a participating school for the purposes of calculating and documenting response rates. However, data from such schools were included in the database and contributed to the estimates included in the initial PISA international report. Data from schools with a student participation rate of less than 25 per cent were not included in the database, and such schools were also regarded as non-respondents.

The rationale for this approach was as follows. There was concern that, in an effort to meet the requirements for school response rates, a national centre might accept participation from schools that would not make a concerted effort to have students attend the assessment sessions. To avoid this, a standard for student participation was required for each individual school, in order that the school be regarded as a participant. This standard was set at 50 per cent. However, in many countries there were a few schools that conducted the assessment without meeting that standard. Thus a judgement was needed to decide if the data from students in such schools should be used in the analyses, given that the students had already been assessed. If the students from such schools were retained, non-response bias would be introduced to the extent that the students who were absent were different in achievement from those who attended the testing session, and such a bias is magnified by the relative sizes of these two groups. If one chose to delete all assessment

Figure 4.1 ■ **School response rate standards**





data from such schools, then non-response bias would be introduced to the extent that the school was different from others in the sample, and sampling variance is increased because of sample size attrition.

The judgement was made that, for a school with between 25 and 50 per cent student response, the latter source of bias and variance was likely to introduce more error into the study estimates than the former, but with the converse judgement for those schools with a student response rate below 25 per cent. Clearly the cut-off of 25 per cent is an arbitrary one, as one would need extensive studies to try to establish this cut-off empirically. However, it is clear that, as the student response rate decreases within a school, the bias from using the assessed students in that school will increase, while the loss in sample size from dropping all of the students in the school will rapidly decrease.

These PISA standards applied to weighted school response rates. The procedures for calculating weighted response rates are presented in Chapter 8. Weighted response rates weight each school by the number of students in the population that are represented by the students sampled from within that school. The weight consists primarily of the enrolment size of 15-year-old students in the school, divided by the selection probability of the school. Because the school samples were in general selected with probability proportional to size, in most countries most schools contributed equal weights, so that weighted and unweighted school response rates were very similar. Exceptions could occur in countries that had explicit strata that were sampled at very different rates. Details as to how the PISA participants performed relative to these school response rate standards are included in Chapter 12 and Chapter 15.

Student response rates

A response rate of 80 per cent of selected students in participating schools was required. A student who had participated in the original or follow-up cognitive sessions was considered to be a participant. A student response rate of 50 per cent within each school was required for a school to be regarded as participating: the overall student response rate was computed using only students from schools with at least a 50 per cent response rate. Again, weighted student response rates were used for assessing this standard. Each student was weighted by the reciprocal of their sample selection probability.

MAIN STUDY SCHOOL SAMPLE

Definition of the national target population

NPMs were first required to confirm their dates of testing and age definition with the PISA consortium. Once these were approved, NPMs were alerted to avoid having the possible drift in the assessment period lead to an unapproved definition of the national target population.

Every NPM was required to define and describe their country's national desired target population and explain how and why it might deviate from the international target population. Any hardships in accomplishing complete coverage were specified, discussed and approved or not, in advance. Where the national desired target population deviated from full national coverage of all eligible students, the deviations were described and enrolment data provided to measure how much that coverage was reduced.

School-level and within-school exclusions from the national desired target population resulted in a national-defined target population corresponding to the population of students recorded on each country's school sampling frame. Schools were usually excluded for practical reasons such as increased survey costs, complexity in the sample design, or difficult test conditions. They could be excluded, depending on the

percentage of 15-year-old students involved, if they were geographically inaccessible (but not part of a region omitted from the national desired target population), or if it was not feasible to administer the PISA assessment. These difficulties were mainly addressed by modifying the sample design to reduce the number of such schools selected, rather than to exclude them, and exclusions from the national desired target population were held to a minimum and were almost always below 0.5 per cent. Schools with students that would all be excluded through the within-school exclusion categories could be excluded up to a maximum of 2 per cent. Otherwise, countries were instructed to include the schools but to administer the PISA UH booklet,[1] consisting of a subset of PISA assessment items deemed more suitable for students with special educational needs.

Within-school, or student-level, exclusions were generally expected to be less than 2.5 per cent in each country, allowing an overall level of exclusion within a country to be no more than 5 per cent. Because definitions of within-school exclusions could vary from country to country, however, NPMs were asked to adapt the following rules to make them workable in their country, but still to code them according to the PISA international coding scheme.

Within participating schools, all eligible students (*i.e.* born within the defined time period, regardless of grade) were to be listed. From this, either a sample of 35 students was randomly selected, or all students were selected if there were fewer than 35 15-year-olds. The lists had to include sampled students deemed to meet one of the categories for exclusion, and a variable maintained to briefly describe the reason for exclusion. This made it possible to estimate the size of the within-school exclusions from the sample data.

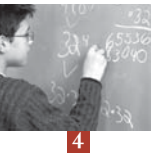
It was understood that the exact extent of within-school exclusions would not be known until the within-school sampling data were returned from participating schools, and sampling weights computed. Country participant projections for within-school exclusions provided before school sampling were known to be estimates.

NPMs were made aware of the distinction between within-school exclusions and non-response. Students who could not take the achievement tests because of a permanent condition were to be excluded and those with a temporary impairment at the time of testing, such as a broken arm, were treated as non-respondents along with other absent sampled students.

Exclusions by country are documented in Chapter 12.

The sampling frame

All NPMs were required to construct a school sampling frame to correspond to their national defined target population. This was defined by the sampling preparation manual[2] as a frame that would provide complete coverage of the national defined target population without being contaminated by incorrect or duplicate entries or entries referring to elements that were not part of the defined target population. Initially, this list was to include any school that could have 15-year-old students, even those who might later be excluded, or deemed ineligible because they had no eligible students at the time of data collection. The quality of the sampling frame directly affects the survey results through the schools' probabilities of selection and therefore their weights and the final survey estimates. NPMs were therefore advised to be very careful in constructing their frames, while realising that the frame depends largely on the availability of appropriate information about schools and students.



All but two countries used school-level sampling frames as their first stage of sample selection. The sampling preparation manual indicated that the quality of sampling frames for both two and three-stage designs would largely depend on the accuracy of the approximate enrolment of 15-year-olds available (*ENR*) for each first-stage sampling unit. A suitable *ENR* value was a critical component of the sampling frames since selection probabilities were based on it for both two and three-stage designs. The best *ENR* for PISA would have been the number of currently enrolled 15-year-old students. Current enrolment data, however, were rarely available at the time of sampling, which meant using alternatives. Most countries used the first-listed available option from these alternatives:

- Student enrolment in the target age category (15-year-olds) from the most recent year of data available;
- If 15-year-olds tend to be enrolled in two or more grades, and the proportions of students who are 15 in each grade are approximately known, the 15-year-old enrolment can be estimated by applying these proportions to the corresponding grade-level enrolments;
- The grade enrolment of the modal grade for 15-year-olds; or
- Total student enrolment, divided by the number of grades in the school.

The sampling preparation manual noted that if reasonable estimates of *ENR* did not exist or if the available enrolment data were too out of date, schools might have to be selected with equal probabilities. This situation occurred for only one country (Greece).

Besides *ENR* values, NPMs were instructed that each school entry on the frame should include at minimum:

- School identification information, such as a unique numerical national identification, and contact information such as name, address and phone number; and
- Coded information about the school, such as region of country, school type and extent of urbanisation, which could be used as stratification variables.[3]

As noted, three-stage designs and area-level sampling frames were used by two countries where a comprehensive national list of schools was not available and could not be constructed without undue burden, or where the procedures for administering the test required that the schools be selected in geographic clusters. As a consequence, area-level sampling frames introduced an additional stage of frame creation and sampling (called the first stage of sampling) before actually sampling schools (the second stage of sampling). Although generalities about three-stage sampling and using an area-level sampling frame were outlined in the sampling preparation manual (for example that there should be at least 80 first-stage units and about half of them needed to be sampled), NPMs were also instructed in the sampling preparation manual that the more detailed procedures outlined there for the general two-stage design could easily be adapted to the three-stage design. NPMs using a three-stage design were also asked to notify the consortium, and received additional support in using an area-level sampling frame. The countries that used a three-stage design were the Russian Federation and Turkey.

Stratification

Prior to sampling, schools were to be ordered, or stratified, on the sampling frame. Stratification consists of classifying schools into like groups according to some variables – referred to as stratification variables.

Stratification in PISA was used to:

- Improve the efficiency of the sample design, thereby making the survey estimates more reliable;
- Apply different sample designs, such as disproportionate sample allocations, to specific groups of schools, such as those in states, provinces, or other regions;
- Make sure that all parts of a population were included in the sample; and
- Ensure adequate representation of specific groups of the target population in the sample.

There were two types of stratification possible: explicit and implicit. Explicit stratification consists of building separate school lists, or sampling frames, according to the set of explicit stratification variables under consideration. Implicit stratification consists essentially of sorting the schools within each explicit stratum by a set of implicit stratification variables. This type of stratification is a very simple way of ensuring a strictly proportional sample allocation of schools across all implicit strata. It can also lead to improved reliability of survey estimates, provided that the implicit stratification variables being considered are correlated with PISA achievement (at the school level). Guidelines were provided on how to go about choosing stratification variables.

Table 4.1 provides the explicit stratification variables used by each country, as well as the number of explicit strata, and the variables and their number of levels used for implicit stratification.[4]

Treatment of small schools in stratification

In PISA, small, moderately small and very small schools were identified, and all others were considered large. A small school had an approximate enrolment of 15-year-olds (*ENR*) below the target cluster size (*TCS* = 35 in most countries) of numbers of students to be sampled from schools with large enrolments. A very small school had an *ENR* less than one-half the *TCS* – 17 or less in most countries. A moderately small school had an *ENR* in the range of *TCS*/*2* to *TCS*. Unless they received special treatment, small schools in the sample could reduce the sample size of students for the national sample to below the desired target because the in-school sample size would fall short of expectations. A sample with many small schools could also be an administrative burden. To minimise these problems, procedures for stratifying and allocating school samples were devised for small schools on the sampling frame.

To determine what was needed – a single stratum of small schools (very small and moderately small combined), a stratum of very small schools only, two strata, one of very small schools and one of moderately small schools, or no small school strata if none of the following conditions were true – the sampling preparation manual stipulated that if:

- The percentage of students in very small schools was 1 per cent or more and the percentage of students in moderately small schools was 4 per cent or more, then an explicit stratum of moderately small schools and an explicit stratum for very small schools was required.
- Otherwise, if the percentage of students in very small schools was 1 per cent or more, a stratum for very small schools was needed, but no stratum for moderately small schools.
- Otherwise, if the percentage of students in very small schools was less than 1 per cent, and the percentage of students in moderately small schools was 4 per cent or more, a combined stratum for small schools, which included all very small and moderately small schools, was needed.

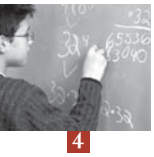
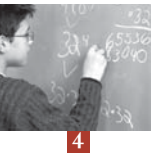


Table 4.1 ■ Stratification variables

Country	Explicit stratification variables	Number of explicit strata	Implicit stratification variables
Australia	State/territory (8) Sector (3) plus 1 for ACT School size (1)	26	Urban/rural (2)
Austria	School type (16) School size (2)	18	District (121)
Belgium			
Belgium (Flanders)	Form of education (5) Public/private (2) School size (2)	11	Index of overaged students
Belgium (French)	Public/private (4) Special education/other (2)	8	School size (3); index of overaged students
Belgium (German)	one explicit stratum (all of German Belgium)	1	School type (3); school size (2)
Brazil	Regions (5) Public/private (2) Size (2)	12	Type of public, for public school strata (2); urban/rural (2); school infrastructure index (4)
Canada	Province (10) Language (3) School size (25) Certainty selections	71	Public/private (2); urban/rural (2)
Czech Republic	School type (6) Regions (14) (only for school types 1 and 2) School size (2)	34	Regions (14) (for schools types 3, 4, 5 and 6)
Denmark	School size (3)	3	Type of school (4); county (15)
Finland	Region (6) Urban/rural (2)	12	None
France	School type (4) School size (2)	6	None
Germany	School category (3) State (16) for normal schools	18	School type for normal schools (5); state for other schools (16)
Greece	Region (10) Public/private (2) Evening schools (1)	13	School type (4); public/private (2) when both in an explicit stratum
Hong Kong-China	School type (3)	3	Student academic intake (3); funding source for independent schools (2)
Hungary	School type (4) Small primary schools excluded from TIMSS (1)	5	Geographical region (8)
Iceland	Geographical region (9)	9	Urban/rural (2); school size (4)
Indonesia	Province (26) School size (2)	28	Type of school (5); public/private (2); national achievement score categories (3)
Ireland	School size (3)	3	School type (3); school gender composition categories (5)
Italy	Geographical region (11) Programme (4) School size (2)	44	Public/private (2)
Japan	Public/private (2) School type (2)	4	Levels of proportion of students taking university/college entrance exams (4)
Latvia	School size (3)	3	Urbanicity (3); school type (3)
Liechtenstein	One explicit stratum (all of Liechtenstein)	1	None
Luxembourg	School type (3)	3	None
Macao-China	School type (3)	3	None

Table 4.1 ■ **Stratification variables** (*continued*)

Country	Explicit stratification variables	Number of explicit strata	Implicit stratification variables
Mexico	State (32) School size (2), certainty selection	52	School type (6); urban/rural (2); school level (3); school programme (4 for lower secondary, 3 for upper secondary)
Netherlands	School track level (2)	2	School type depending on track (6)
New Zealand	Certainty/non-certainty (2)	2	Public/private (2); socio-economic status category (3) and urban/rural (2) for public schools
Norway	School type (2), size (4)	4	None
Poland	School type (2)	2	Urbanicity (4)
Portugal	Geographical area (7) School size (2)	9	Public/private (2); socio-economic status category (4)
Republic of Korea	School type (3) Urbanisation (3) School size (2)	10	School level (2)
Russian Federation	PSU (45)	45	School type (3); urbanicity (5)
Serbia	Geographic region (8) Certainty selections	10	Urban/rural (2); school type (7); Hungarian students or not (2)
Slovak Republic	Primary/secondary (2) Region (8) school size (2)	20	School type (9); language (2); authority (9)
Spain	Region (17) Public/private (2) Teaching modality for Basque (3) School size (2)	46	Size of town for Catalonia (3); postal code (provinces and districts) for all
Sweden	Public/private (2) School size (2) Urbanicity (5) Upper secondary	9	Private: upper secondary or not (2); geographical area (22); urbanicity (9); public: school type (2); responsible authority (2); geographical area (22), income quartile (4)
Switzerland	Language (3) Canton/region (7) School has grade 9 or not (2) Public/private (2) Certainty selections	37	School type (29); Canton (26) in strata where several Cantons
Thailand	Department (8) School level (3) School size (2)	15	Region (13)
Tunisia	Geographical area (2)	2	Levels of grade repeating for three school levels
Turkey	PSU (40) School size (1) Certainty selections	44	School type (18)
United Kingdom			
England	School size (2)	2	School type (2); exam result categories for not small schools (7); gender mix for independent, non-small schools (3); LEA (150)
Northern Ireland	Certainty/non-certainty (2)	2	School type (3); exam results for secondary and grammar (4 and 3 levels respectively); region (5)
Scotland	School S-grade attainment (5)	5	None
Wales	One explicit stratum (all of Wales)	1	Secondary/independent (2); exam result categories (4) for secondary
United States	One explicit stratum (all of the United States)	1	Grade span (5); public/private (2); region of country (4); urbanicity area (8); minority status (2)
Uruguay	School type (4) Area (3) School size (2)	12	Programme (3 or 7 depending on school type); shift (4 or 5 depending on school type); area (3) for private schools



The small school strata were always sorted first by the explicit stratum to which they originally belonged, followed by the other defined implicit stratification variables.

When small schools were explicitly stratified, it was important to ensure that an adequate sample was selected without selecting too many small schools as this would lead to too few students in the assessment. In this case, the entire school sample would have to be increased to meet the target student sample size.

The sample had to be proportional to the number of students and not to the number of schools. Suppose that 10 per cent of students attend moderately small schools, 10 per cent very small schools and the remaining 80 per cent attend large schools. In the sample of 5 250, 4 200 students would be expected to come from large schools (*i.e.* 120 schools with 35 students), 525 students from moderately small schools and 525 students from very small schools. If moderately small schools had an average of 25 students, then it would be necessary to include 21 moderately small schools in the sample. If the average size of very small schools was 10 students, then 52 very small schools would be needed in the sample and the school sample size would be equal to 193 schools rather than 150.

To balance the two objectives of selecting an adequate sample of explicitly stratified small schools, a procedure was recommended that assumes identifying strata of both very small and moderately small schools. The underlying idea is to under-sample by a factor of two the very small school stratum and to increase proportionally the size of the large school strata. When there was just a single small school stratum, the procedure was modified by ignoring the parts concerning very small schools. The formulae below also assume a school sample size of 150 and a student sample size of 5 250.

- *Step 1:* From the complete sampling frame, find the proportions of total *ENR* that come from very small schools (*P*), moderately small schools (*Q*), and larger schools (those with *ENR* of at least *TCS*) (*R*). Thus, $P + Q + R = 1$.
- *Step 2:* Calculate the figure *L*, where $L = 1 + (P/2)$. Thus *L* is a positive number slightly more than 1.0.
- *Step 3:* The minimum sample size for larger schools is equal to $150 \times R \times L$, rounded to the nearest integer. It may need to be enlarged because of national considerations, such as the need to achieve minimum sample sizes for geographic regions or certain school types.
- *Step 4:* Calculate the mean value of *ENR* for moderately small schools (*MENR*), and for very small schools (*VENR*). *MENR* is a number in the range of *TCS*/2 to *TCS*, and *VENR* is a number no greater than *TCS*/2.
- *Step 5:* The number of schools that must be sampled from the stratum of moderately small schools is given by: $(5\,250 \times Q \times L)/(MENR)$.
- *Step 6:* The number of schools that must be sampled from the stratum of very small schools is given by: $(2\,625 \times P \times L)/(VENR)$.

To illustrate the steps, suppose that in participant country *X*, the *TCS* is equal to 35, with 0.1 of the total enrolment of 15-year-olds each in moderately small schools and in very small schools. Suppose that the average enrolment in moderately small schools is 25 students, and in very small schools it is 10 students. Thus $P = 0.1$, $Q = 0.1$, $R = 0.8$, $MENR = 25$ and $VENR = 10$.

From *Step 2*, $L = 1.05$. Then (*Step 3*) the sample size of larger schools must be at least $150 \times (0.80 \times 1.05) = 126.3$. That is, at least 126 of the larger schools must be sampled. From *Step 5*, the number of moderately small schools required is $(5\,250 \text{ x } 0.1 \times 1.05)/25 = 22.1$ – that is, 22 schools. From *Step 6*, the number of very small schools required is $(2\,625 \times 0.1 \times 1.05)/10 = 27.6$ – that is, 28 schools.

This gives a total sample size of 126 + 22 + 28 = 176 schools, rather than just 150, or 193 as calculated above. Before considering school and student non-response, the larger schools will yield a sample of 126 × 35 = 4 410 students. The moderately small schools will give an initial sample of approximately 22 × 25 = 550 students, and very small schools will give an initial sample size of approximately 28 × 10 = 280 students. The total initial sample size of students is therefore 4 410 + 550 + 280 = 5 240.

Assigning a measure of size to each school

For the probability proportional-to-size sampling method used for PISA, a measure of size (*MOS*) derived from *ENR* was established for each school on the sampling frame. Where no explicit stratification of very small schools was required *or* if small schools (including very small schools) were separately stratified because school size was an explicit stratification variable and they did not account for 5 per cent or more of the target population, *MOS* was constructed as: *MOS* = max (*ENR, TCS*).

The measure of size was therefore equal to the enrolment estimate, unless it was less than the *TCS*, in which case it was set equal to the target cluster size. In most countries, *TCS* = 35 so that the *MOS* was equal to *ENR* or 35, whichever was larger.

As sample schools were selected according to their size (PPS), setting the measure of size of small schools to 35 is equivalent to drawing a simple random sample of small schools.

School sample selection

Sorting the sampling frame

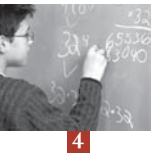
The sampling preparation manual indicated that, prior to selecting schools from the school sampling frame, schools in each explicit stratum were to be sorted by variables chosen for implicit stratification and finally by the *ENR* value within each implicit stratum. The schools were first to be sorted by the first implicit stratification variable, then by the second implicit stratification variable within the levels of the first sorting variable, and so on, until all implicit stratification variables were exhausted. This gave a cross-classification structure of cells, where each cell represented one implicit stratum on the school sampling frame. The sort order was alternated between implicit strata, from high to low and then low to high, etc., through all implicit strata within an explicit stratum.

School sample allocation over explicit strata

The total number of schools to be sampled in each country needed to be allocated among the explicit strata so that the expected proportion of students in the sample from each explicit stratum was approximately the same as the population proportions of eligible students in each corresponding explicit stratum. There were two exceptions. If an explicit stratum of very small schools was required, students in them had smaller percentages in the sample than those in the population. To compensate for the resulting loss of sample, the large school strata had slightly higher percentages in the sample than the corresponding population percentages. The other exception occurred if only one school was allocated to any explicit stratum. In this case, two schools were allocated for selection in the stratum to aid with variance estimation.

Determining which schools to sample

The PPS systematic sampling method used in PISA first required the computation of a sampling interval for each explicit stratum. This calculation involved the following steps:



- Recording the total measure of size, *S,* for all schools in the sampling frame for each specified explicit stratum;
- Recording the number of schools, *D*, to be sampled from the specified explicit stratum, which was the number allocated to the explicit stratum;
- Calculating the sampling interval, *I*, as follows: $I = S/D$; and
- Recording the sampling interval, *I*, to four decimal places.

Next, a random number (drawn from a uniform distribution) had to be selected for each explicit stratum**.** The generated random number (*RN*) was to be a number between 0 and 1 and was to be recorded to four decimal places. The next step in the PPS selection method in each explicit stratum was to calculate selection numbers – one for each of the *D* schools to be selected in the explicit stratum. Selection numbers were obtained using the following method:

- Obtaining the first selection number by multiplying the sampling interval, *I*, by the random number, *RN*. This first selection number was used to identify the first sampled school in the specified explicit stratum;
- Obtaining the second selection number by simply adding the sampling interval, *I*, to the first selection number. The second selection number was used to identify the second sampled school; and
- Continuing to add the sampling interval, *I*, to the previous selection number to obtain the next selection number. This was done until all specified line numbers (1 through *D*) had been assigned a selection number.

Thus, the first selection number in an explicit stratum was $RN \times I$, the second selection number was $(RN \times I) + I$, the third selection number was $(RN \times I) + I + I$, and so on.

Selection numbers were generated independently for each explicit stratum, with a new random number selected for each explicit stratum.

PISA/TIMSS overlap control

Because the main study for PISA 2003 and the 2003 Trends in International Mathematics and Science Study (TIMSS) would occur at approximately the same time, an overlap control procedure was used for countries (Belgium Flanders, Spain, Sweden, Australia, Scotland, the Netherlands and Tunisia) who wished for there to be a minimum (or a maximum) of the same schools to be sampled for each study. This procedure could only be done if the same school identifiers were used on the TIMSS and PISA school frames and if the schools used on each frame were the same.

The TIMSS samples were usually selected before the PISA samples. Thus, for countries requesting overlap control, the TIMSS International Study Center supplied the PISA consortium with their school frames, with the school IDs, the school probability of selection, and an indicator showing which schools had been sampled for TIMSS. Only in two countries where overlap control was requested (the Netherlands and Scotland) did PISA select school samples first. In these cases, schools were sampled as usual unless there were any PISA school probabilities of selection greater than 0.5 (see discussion below).

TIMSS and PISA sample selections could generally avoid overlap of schools if any schools which would have been selected with high probability for either study had their probabilities capped at 0.5. Such an action

would make each study's sample slightly less than optimal, but this was deemed acceptable when weighed against the possibility of low response rates due to school burden. Each study's NPM had to decide if this was the path they wished to adopt. If they decided against this capping of probabilities, then it was possible for some large schools to be in both the TIMSS and PISA samples.

To control overlap, the sample selection of schools for PISA adopted a modification of the approach due to Keyfitz (1951), based on the Bayes Theorem.

Suppose that *PROBT* is the TIMSS probability of selection, and *PROBP* is the usual PISA probability of selection, then a conditional probability of selection into PISA, *CPROB*, is determined, based upon whether there is a desire to minimise or maximise the overlap between the TIMSS and PISA samples.

If the desire is to minimise the overlap then *CPROB* is defined as follows:

$$CPROB = \begin{cases} \max\left[0, \left(\dfrac{PROBT + PROBP - 1}{PROBT}\right)\right] & \text{if the school was TIMSS selected} \\ \min\left[1, \dfrac{PROBP}{(1 - PROBT)}\right] & \text{if the school was not TIMSS selected} \\ PROBP & \text{if the school was not a TIMSS eligible school} \end{cases} \quad (4.1)$$

If the desire is to maximise the overlap then *CPROB* is defined as follows:

$$CPROB = \begin{cases} \min\left[1, \left(\dfrac{PROBP}{PROBT}\right)\right] & \text{if the school was TIMSS selected} \\ \max\left[0, \dfrac{(PROBP - PROBT)}{(1 - PROBT)}\right] & \text{if the school was not TIMSS selected} \\ PROBP & \text{if the school was not a TIMSS eligible school} \end{cases} \quad (4.2)$$

Then a conditional *MOS* variable was created to coincide with these conditional probabilities as follows:

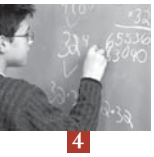
CMOS = *CPROB* × stratum sampling interval (rounded to 4 decimal places).

The PISA school sample was then selected using the line numbers created as usual (see below), but applied to the cumulated *CMOS* values (as opposed to the cumulated *MOS* values). Note that it was possible that the resulting PISA sample size could be a bit lower or higher than the originally assigned sample size, but this was deemed acceptable.

Identifying the sampled schools

The next task was to compile a cumulative measure of size in each explicit stratum of the school sampling frame that determined which schools were to be sampled. Sampled schools were identified as follows.

Let Z denote the first selection number for a particular explicit stratum. It was necessary to find the first school in the sampling frame where the cumulative *MOS* equalled or exceeded Z. This was the first sampled school. In other words, if C_s was the cumulative *MOS* of a particular school S in the



sampling frame and $C_{(s-1)}$ was the cumulative *MOS* of the school immediately preceding it, then the school in question was selected if: C_s was greater than or equal to Z, and $C_{(s-1)}$ was strictly less than Z. Applying this rule to all selection numbers for a given explicit stratum generated the original sample of schools for that stratum.

Identifying replacement schools

Each sampled school in the main survey was assigned two replacement schools from the sampling frame, identified as follows. For each sampled school, the schools immediately preceding and following it in the explicit stratum were designated as its replacement schools. The school immediately following the sampled school was designated as the first replacement and labelled R_1, while the school immediately preceding the sampled school was designated as the second replacement and labelled R_2. The *Sampling Preparation Manual* noted that in small countries, there could be problems when trying to identify two replacement schools for each sampled school. In such cases, a replacement school was allowed to be the potential replacement for two sampled schools (a first replacement for the preceding school, and a second replacement for the following school), but an actual replacement for only one school. Additionally, it may have been difficult to assign replacement schools for some very large sampled schools because the sampled schools appeared very close to each other in the sampling frame. There were times when it was only possible to assign a single replacement school, or even none, when two consecutive schools in the sampling frame were sampled.

Exceptions were allowed if a sampled school happened to be the first or last school listed in an explicit stratum. In these cases the two schools immediately following or preceding it were designated as replacement schools

Assigning school identifiers

To keep track of sampled and replacement schools in the PISA database, each was assigned a unique, three-digit school code and two-digit stratum code (corresponding to the explicit strata) sequentially numbered starting with one within each explicit stratum. For example, if 150 schools are sampled from a single explicit stratum, they are assigned identifiers from 001 to 150. First replacement schools in the main survey are assigned the school identifier of their corresponding sampled schools, incremented by 300. For example, the first replacement school for sampled school 023 is assigned school identifier 323. Second replacement schools in the main survey are assigned the school identifier of their corresponding sampled schools, but incremented by 600. For example, the second replacement school for sampled school 136 took the school identifier 636.

Tracking sampled schools

NPMs were encouraged to make every effort to confirm the participation of as many sampled schools as possible to minimise the potential for non-response biases. They contacted replacement schools after all contacts with sampled schools were made. Each sampled school that did not participate was replaced if possible. If both an original school and a replacement participated, only the data from the original school were included in the weighted data provided that at least 50% of the eligible, non-excluded students had participated. If this was not the case, it was permissible for the original school to be labelled as a non-respondent and the replacement school as the respondent, provided that the replacement school had at least 50% of the eligible, non-excluded students as participants.

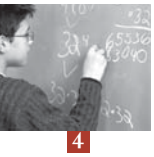
© OECD 2005

Monitoring school sampling

For PISA 2003, it was a strong recommendation that the consortium select the school samples. This was incorporated into the 2003 procedures to alleviate the weighting difficulties caused by receiving school frame files in many different formats. Only Finland, France, Germany, Japan, Poland and the United States selected their own school samples, for reasons varying from timing conflicts, to confidentiality restraints, to having complex designs because of planned national option sampling or internal overlap control with other surveys. The consortium checked all samples in detail. All countries were required to submit sampling forms 1 (time of testing and age definition), 2 (national desired target population), 3 (national defined target population), 4 (sampling frame description), 5 (excluded schools), 7 (stratification) and 11 (school sampling frame). The consortium completed and returned the others (forms 6, 8, 9, 10 and 12) for countries for which they did the sampling. Otherwise, the country also submitted these other forms for approval. Table 4.2 provides a summary of the information required on each form and the timetables, which depended on national assessment periods. See Appendix 1 for copies of the sampling forms.

Table 4.2 ■ **Schedule of school sampling activities**

Activity	Submit to Consortium	Due date
Specify time of testing and age definition of population to be tested	Sampling form 1 - Time of testing and age definition	Submit at least six months before the beginning of testing
Define national desired target population	Sampling form 2 - National desired target population	Submit at least six months before the beginning of testing
Define national defined target population	Sampling form 3 - National defined target population	Submit at least six months before the beginning of testing
Create and describe sampling frame	Sampling form 4 - Sampling frame description	Submit at least five months before the beginning of testing
Decide on schools to be excluded from sampling frame	Sampling form 5 - Excluded schools	Submit at least five months before the beginning of testing
Decide how to treat small schools	Sampling form 6 - Treatment of small schools	The consortium will complete and return this form to the NPM about four months before the beginning of testing
Decide on explicit and implicit stratification variables	Sampling form 7 - Stratification	Submit at least five months before the beginning of testing
Describe population within strata	Sampling form 8 - Population counts by strata	The consortium will complete and return this form to the NPM about three months before the beginning of testing
Allocate sample over explicit strata	Sampling form 9 - Sample allocation by explicit strata	The consortium will complete and return this form to the NPM about three months before the beginning of testing
Select the school sample	Sampling form 10 - School sample selection	The consortium will complete and return this form to the NPM about three months before the beginning of testing
Identify sampled schools, replacement schools and assign PISA school IDs	Sampling form 11 - School sampling frame	Submit five months before the beginning of testing. The consortium will return this form to the NPM with sampled schools and their replacement schools identified and with PISA IDS assigned about three months before the beginning of testing.
Create a school tracking form	Sampling form 12 - School tracking form	Submit within one month of the end of the data collection period



Once received from each country, each form was reviewed and feedback was provided to the country. Forms were only approved after all criteria were met. Approval of deviations was only given after discussion and agreement by the consortium. In cases where approval could not be granted, countries were asked to make revisions to their sample design and sampling forms.

Checks that were performed in the monitoring of each form follow. All entries were observed in their own right but those below are additional matters explicitly examined.

Sampling form 1: Time of testing and age definition

- Assessment dates had to be appropriate for the selected target population dates.
- Assessment dates could not cover more than a 42-day period.

Sampling form 2: National desired target population

- Large deviations between the total national number of 15-year-olds and the enrolled number of 15-year-olds were questioned.
- Large increases or decreases in population numbers compared to those from PISA 2000 were queried.
- Any population to be omitted from the international desired population was noted and discussed, especially if the percentage of 15-year-olds to be excluded was more than 2 per cent.
- Calculations were verified.
- For any countries using a three-stage design, a sampling form 2 also needed to be completed for the full national desired population as well as for the population in the sampled regions.

Sampling form 3: National defined target population

- The population figure in the first question needed to correspond with the final population figure on sampling form 2.
- Reasons for excluding schools were checked for appropriateness.
- The number and percentage of students to be excluded at the school level and whether the percentage was less than the maximum percentage allowed for such exclusions, were checked.
- Calculations were verified and the overall coverage figures were assessed.
- For any countries using a three-stage design, sampling form 3 also needed to be completed for the full national defined population as well as for the population in the sampled regions.

Sampling form 4: Sampling frame description

- Special attention was paid to countries who reported on this form that a three-stage sampling design was to be implemented, and additional information was sought from countries in such cases to ensure that the first-stage sampling was done adequately.
- The type of school-level enrolment estimate and the year of data availability were assessed for reasonableness.

Sampling form 5: Excluded schools

- The number of schools and the total enrolment figures, as well as the reasons for exclusion, were checked to ensure correspondence with figures reported on sampling form 3 about school-level exclusions.

© OECD 2005

Sampling form 6: Treatment of small schools

- Calculations were verified, as was the decision about whether or not a moderately small schools stratum and/or a very small schools stratum were needed.

Sampling form 7: Stratification

- Since explicit strata are formed to group like schools together to reduce sampling variance and to ensure appropriate representativeness of students in various school types, using variables that might have an effect on outcomes, each country's choice of explicit stratification variables was assessed. If a country was known to have school tracking, and tracks or school programmes were not among the explicit stratifiers, a suggestion was made to include this type of variable.
- If no implicit stratification variables were noted, suggestions were made about ones that might be used.
- The sampling frame was checked to ensure that the stratification variables were available for all schools. Different explicit strata were allowed to have different implicit stratifiers.

Sampling form 8: Population counts by strata

- Counts on sampling form 8 were compared to counts arising from the frame. Any differences were queried and almost always corrected.

Sampling form 9: Sample allocation by explicit strata

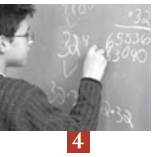
- All explicit strata had to be accounted for on sampling form 9.
- All explicit strata population entries were compared to those determined from the sampling frame.
- The calculations for school allocation were checked to ensure that schools were allocated to explicit strata based on explicit stratum student percentages and not explicit stratum school percentages.
- The percentage of students in the sample for each explicit stratum had to be close to the percentage in the population for each stratum (very small schools strata were an exception since under-sampling was allowed).
- The overall number of schools to be sampled was checked to ensure that at least 150 schools would be sampled.
- The overall number of students to be sampled was checked to ensure that at least 5 250 students would be sampled.

Sampling form 10: School sample selection

- All calculations were verified.
- Particular attention was paid to the four decimal places that were required for both the sampling interval and the random number.

Sampling form 11: School sampling frame

- The frame was checked for proper sorting according to the implicit stratification scheme and enrolment values, and the proper assignment of the measure of size value, especially for moderately small and very small schools. The accumulation of the measure of size values was also checked for each explicit stratum. This final cumulated measure of size value for each stratum had to correspond to the total measure of size value on sampling form 10 for each explicit stratum. Additionally, each line selection number was checked against the frame cumulative measure of size figures to ensure that the correct schools were sampled. Finally, the assignment



of replacement schools and PISA identification numbers were checked to ensure that all rules laid out in the sampling manual were adhered to. Any deviations were discussed with each country and either corrected or the deviations accepted.

Sampling form 12: School tracking form

- Sampling form 12 was checked to see that the PISA identification numbers on this form matched those on the sampling frame.
- Checks were made to ensure that all sampled and replacement schools were accounted for.
- Checks were also made to ensure that status entries were in the requested format.

Student samples

Student selection procedures in the main study were the same as those used in the field trial. Student sampling was generally undertaken at the national centres using the consortium software, *KeyQuest*, from lists of all eligible students in each school that had agreed to participate. These lists could have been prepared at national, regional, or local levels as data files, computer-generated listings, or by hand, depending on who had the most accurate information. Since it was very important that the student sample be selected from accurate, complete lists, the lists needed to be prepared not too far in advance of the testing and had to list all eligible students. It was suggested that the lists be received one to two months before testing so that the NPM would have the time to select the student samples.

Some countries chose student samples that included students aged 15 and/or enrolled in a specific grade (*e.g.* grade 10). Thus, a larger overall sample, including 15-year-old students and students in the designated grade (who may or may not have been aged 15) were selected. The necessary steps in selecting larger samples are highlighted where appropriate in the following steps. Only Iceland, the Czech Republic, and Switzerland selected grade samples, and only the Czech Republic used the standard method described here. For Iceland, the sample was called a grade sample because over 99.5 per cent of the PISA eligible 15-year-olds were in the grade sampled. Switzerland supplemented the standard method with an additional sample of grade-eligible students which was selected by first selecting grade 9 classes within PISA sampled schools that had this grade.

Preparing a list of age-eligible students

Appendix 17 shows an example of the student listing form, as well as school instructions about how to prepare the lists. Each school drawing an additional grade sample was to prepare a list of age and grade-eligible students that included all students in the designated grade (*e.g.* grade 10); and all other 15-year-old students (using the appropriate 12-month age span agreed upon for each country) currently enrolled in other grades. NPMs were to use the student listing form as shown in the Appendix 17 example but could develop their own instructions. The following were considered important:

- Age-eligible students were all students born in 1987 (or the appropriate 12-month age span agreed upon for the country).
- The list was to include students who might not be tested due to a disability or limited language proficiency.
- Students who could not be tested were to be excluded from the assessment after the student sample was selected.

- It was suggested that schools retain a copy of the list in case the NPM had to call the school with questions.
- A computer list was to be up-to-date at the time of sampling rather than prepared at the beginning of the school year. Students were identified by their unique student identification numbers.

Selecting the student sample

Once NPMs received the list of eligible students from a school, the student sample was to be selected and the list of selected students (*i.e.* the student tracking form) returned to the school. NPMs were encouraged to use *KeyQuest*, the PISA sampling software, to select the student samples.

Preparing instructions for excluding students

PISA was a timed assessment administered in the instructional language(s) of each country and designed to be as inclusive as possible. For students with limited language proficiency or with physical, mental, or emotional disabilities who could not participate, PISA developed instructions in cases of doubt about whether a selected student should be assessed. NPMs used the guidelines given to develop instructions;

Figure 4.2 ■ **Instructions for excluding students**

The following guidelines define general categories for the exclusion of students within schools. These guidelines need to be carefully implemented within the context of each educational system. The numbers to the left are codes to be entered in column (7) of the student tracking form to identify excluded students.

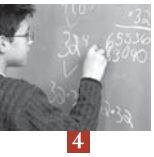
1 *Functionally disabled students:* These are students who are permanently physically disabled in such a way that they cannot perform in the PISA testing situation. Functionally disabled students who can respond to the test should be included in the testing.

2 *Intellectually disabled students:* These are students who are considered in the professional opinion of the school principal or by other qualified staff member to be intellectually disabled or who have been psychologically tested as such. This includes students who are emotionally or mentally unable to follow even the general instructions of the test. However, students should not be excluded solely because of poor academic performance or disciplinary problems.

3 *Students with limited proficiency in the test language:* These are students who are unable to read or speak the language of the test and would be unable to overcome the language barrier in the test situation. Typically, a student who has received less than one year of instruction in the language of the test should be excluded, but this definition may need to be adapted in different countries.

4 *Other,* to be defined as a single reason for exclusion by the NPM before data collection and to be agreed upon with the consortium.

It is important that these criteria be followed strictly for the study to be comparable within and across countries. When in doubt, the student should be included.



School co-ordinators and test administrators needed precise instructions for exclusions (Figure 4.2). The national operational definitions for within-school exclusions were to be well documented and submitted to the consortium for review before testing.

Sending the student tracking form to the school co-ordinator and test administrator

The school co-ordinator needed to know which students were sampled in order to notify them and their teachers (and parents), to update information and to identify the students to be excluded. The student tracking form and guidelines for excluding students were therefore sent about two weeks before the assessment session. It was recommended that a copy of the tracking form be made and kept at the national centre. Another recommendation was to have the NPM send a copy of the form to the test administrator with the assessment booklets and questionnaires in case the school copy was misplaced before the assessment day. The test administrator and school co-ordinator manuals (see Chapter 6) both assumed that each would have a copy.

Notes

1 The UH booklet is described in the section on test design in Chapter 2.

2 Available as the study working document: *SchSampling_Eng1.pdf.*

3 Variables used for dividing the population into mutually exclusive groups so as to improve the precision of sample-based estimates.

4 As countries were requested to sort the sampling frame by school size, school size was also an implicit stratification variable, though it is not listed in Table 17. A variable used for stratification purposes is not necessarily included in the PISA data files.

5

Translation and Cultural Appropriateness of the Test and Survey Material

INTRODUCTION

Translation errors are known to be a major cause of items functioning poorly in international tests. Translation errors are much more frequent than other problems, such as clearly identified discrepancies due to cultural biases or curricular differences.

If a survey is done merely to rank countries or students, this problem can be avoided somewhat, since once the most unstable items have been identified and dropped, the few remaining problematic items are unlikely to affect the overall estimate of a country's mean in any significant way.

The aim of PISA, however, is to develop descriptive scales, and in this case translation errors are of greater concern. The interpretation of a scale can be severely biased by unstable item characteristics from one country to another. PISA has therefore implemented stricter verification procedures for translation equivalence than those used in prior surveys. These procedures included:

- Providing two parallel source versions of the instruments (in English and French), and recommending that each country develop two independent versions in their instruction language (one from each source language), then reconcile them into one national version;
- Appending in the materials frequent translation notes to help with possible translation or adaptation problems;
- Developing detailed translation and adaptation guidelines for the test material, and for revising it after the field trial, as an important part of the PISA national project manager manual;
- Training key staff from each national team on recommended translation procedures; and
- Appointing and training a group of international verifiers (professional translators proficient in English and French, with native command of each target language), to verify the national versions against the source versions.

DOUBLE TRANSLATION FROM TWO SOURCE LANGUAGES

A back translation procedure has long been the most frequently used to ensure linguistic equivalence of test instruments in international surveys. It requires translating the source version of the test (generally English) into the national languages, then translating them back to English and comparing them with the source version to identify possible discrepancies.

A double translation procedure (*i.e.* two independent translations from the source language, with reconciliation by a third person) offers two significant advantages in comparison with the back translation procedure:

- Equivalence of the source and target languages is obtained by using three different people (two translators and a reconciler) who all work on the source and the target versions. In the back translation procedure, by contrast, the first translator is the only one to focus simultaneously on the source and target versions.
- Discrepancies are recorded directly in the target language instead of in the source language, as would be the case in a back translation procedure.

A double translation procedure was used in the Third International Mathematics and Science Study–Repeat (TIMSS–R) instead of the back translation procedure used in earlier studies by the International Association for the Evaluation of Educational Achievement (IEA).

PISA used double translation from two different languages because both back translation and double translation procedures fall short, in that the equivalence of the various national versions depends exclusively on their consistency with a single source version (in general, English). In particular, one would wish for as purely a semantic equivalence as possible as the principle is to measure access that students from different countries have to the same meaning, through written material presented in different languages. However, using a single reference language is likely to give more importance than would be desirable to the formal characteristics of that language. If a single source language is used, its lexical and syntactic features, stylistic conventions, and typical organisational patterns of ideas within the sentence will have more impact than desirable on the target language versions.

Some interesting findings in this respect were reported in the IEA/Reading Comprehension survey (Thorndike, 1973), which showed a better item coherence (factorial structure of the tests, distribution of the discrimination coefficients) between English-speaking countries than across other participating countries.

Resorting to two different languages helps, to a certain extent, tone down problems linked to the impact of cultural characteristics of a single source language. Admittedly, both languages used in PISA share an Indo-European origin, which may be regrettable in this particular case. However, they do represent sets of relatively different cultural traditions, and are both spoken in several countries with different geographic locations, traditions, social structures and cultures.

Other expected advantages from using two source languages in the PISA assessment included:

- The verification of the equivalence between the source and the target versions was performed by four different people who all worked directly on the texts in the relevant national languages (*i.e.* two translators, one national reconciler and the consortium's verifier).
- Many translation problems are due to idiosyncrasies: words, idioms, or syntactic structures in one language appear untranslatable into a target language. In many cases, the opportunity to consult the other source version provided hints at solutions.
- Translation problems can be caused when it is not known how much the source text can be modified. A translation that is too faithful may appear awkward; if it is too free or too literary it is very likely to fail to be equivalent. Having two source versions in different languages (for which the fidelity of the translation has been carefully calibrated and approved by consortium experts) provides benchmarks for a national reconciler that are far more accurate in this respect, and that neither back translation nor double translation from a single language could provide.

Since PISA was the first major international survey using two different source languages, empirical evidence from the PISA 2000 field trial results was collected to explore the consequences of using alternative reference languages in the development phase of the various national versions of the survey materials. The outcomes of this study were reported in Chapter 5 of the *PISA 2000 Technical Report* (OECD, 2002; see also Grisay, 2003). The analyses focused on two issues:

- Could the English and French materials provided by the consortium be considered sufficiently equivalent for use as alternative source versions in the recommended translation procedure?
- Did the recommended procedure actually result in better national versions than other procedures used by some of the participating countries?

As regards the first issue, a systematic comparison of the English and French versions of the texts used as stimuli in the PISA 2000 reading materials was conducted, using readability formulae to assess their relative difficulty in both languages, in terms of lexical and syntactical complexity.

Reasonably high consistency was found; that is, English texts with more abstract or more technical vocabulary, or with longer and more complex sentences, tended to show the same characteristics when translated into French.

Globally, however, the average length of words and of sentences tended to be greater in French than in English, resulting in an increase of about 19 per cent of the character count in the French reading stimuli. When comparing the percentages of correct answers given by English-speaking and French-speaking students to test items associated with stimuli that have either a small or a large increase of length in the French version, a very modest, but statistically significant interaction was found, indicating that the longer French units tended to be (proportionally) more difficult for French-speaking students than those with only slightly greater word count.

This pattern suggested that the additional burden to the reading tasks in countries using the longer version did not seem to be substantial, but the hypothesis of some effect of length on the students' performance could not be discarded.

The PISA 2000 field trial data from the English- and French-speaking participating countries were also explored for possible differences in the number and distribution of items that appeared to have poor psychometric characteristics. No difference was found in the percentages of flawed items (English: 7.5 per cent, French 7.7 per cent; $F=0.05$, $p>0.83$). Over a total of 531 reading, mathematics and science items used in the PISA 2000 field trial, only one reading item was found to be flawed in all four French-speaking countries or communities, but in none of the English-speaking countries or communities. Three other reading items were flawed in three of four French-speaking countries or communities, but in none of the English-speaking countries or communities. Conversely, only four items had flaws in all six English-speaking countries or communities or in four or five of them, but only in one French-speaking country or community. None of the science or mathematics items showed the same kind of imbalance.

As regards the second issue, comparisons were conducted in PISA 2000 between the field trial item statistics from groups of countries that had developed their national versions through the recommended procedure (*i.e.* double translation and reconciliation from both languages) versus those that had used alternative procedures (*e.g.* double translation and reconciliation from the English source only, with or without cross-checks against the French source version), or non-standard procedures (*e.g.* single translation from the English source).

Double translation from English and French appeared to have produced national versions that did not differ significantly in terms of the number of flaws from the versions derived through adaptation from one of the source languages. Double translation from the English source only also appeared to be effective, but only when accompanied by extensive crosschecks against the French source. The average number of flaws was significantly higher in groups of countries that did not use both sources, *i.e.* double translation from one language with no crosschecks against the second source version, or single translation. Single translation proved to be the least trustworthy method.

Due to these positive results, a double translation and reconciliation procedure using both source languages was again recommended in PISA 2003, and countries that had used non-standard procedures in PISA 2000 were encouraged to upgrade their translation methods.

DEVELOPMENT OF SOURCE VERSIONS

Some of the new test materials used in PISA 3003 were prepared by the consortium test development teams on the basis of the submissions received from the participating countries. Items were submitted by 15 different countries, either in their national language or in English. The balance of the materials was prepared by the test development teams themselves in the Netherlands (CITO), Japan (NIER) and Australia (ACER). Then, all materials were circulated, in English, for comments and feedback to the expert groups and the National Project Managers (NPMs).

The French version was developed at this early stage through double translation and reconciliation into French of the English materials, so that any comments from the translation team could be used, along with the comments received from the Expert Groups and the NPMs, in the finalisation of both source versions.

As already shown during the development of the PISA 2000 materials, the translation process proved to be very effective in detecting residual errors overlooked by the test developers, and in anticipating potential translation problems. In particular, a number of ambiguities or pitfall expressions could be spotted and avoided from the beginning by slightly modifying both the English and French source versions; the list of aspects requiring national adaptations could be refined; and further translation notes could be added when needed. In this respect, the development of the French source version served as a translation trial, and probably helped provide NPMs with source material that was somewhat easier to translate or contained fewer potential translation traps than it would have had if a single source had been developed.

The final French source version was reviewed by a French domain expert,[1] for appropriateness of the mathematics and science terminology, and by a native professional French proofreader for linguistic correctness. In addition, an independent verification of the equivalence between the final English and French versions was performed by one of the bilingual English/French translators appointed and trained by the consortium for the international verification of the PISA materials, who used the same procedures and verification checklists as for the verification of all other national versions.

Finally, analyses on possible systematic translation errors that might have occurred in all or most of the national versions adapted from the French source version were conducted, using the main study item statistics from the five French-speaking countries participating in PISA 2003.

PISA TRANSLATION GUIDELINES

The translation guidelines developed in PISA 2000 were revised to include more detailed advice on translation and adaptation of mathematics materials, and additional warnings about common translation errors identified during the verification of the PISA 2000 materials. The guidelines were circulated to the NPMs as part of the PISA 2003 national project manager's manual, and included:

- *PISA requirements in terms of necessary national version(s).* PISA takes as a general principle that students should be tested in the language of instruction used in their school. Therefore, the NPMs of multilingual countries were requested to develop as many versions of the test instruments as there were languages of instruction used in the schools included in their national sample. Cases of minority languages used in only a very limited number of schools could be discussed with the sampling referee to decide whether such schools could be excluded from the target population without affecting the overall quality of the data collection.

- *Information on which parts of the materials had to be double translated, or could be single translated.* Double-translation was required for the tests, questionnaires and for the optional Information Communication Technology and Educational Career instruments, but not for the manuals and other logistic material.
- *Instructions related to the recruitment and training of translators and reconcilers.* The scientific and technical support to be provided to the translation team was also outlined. It was suggested, in particular, that translated material and national adaptations deemed necessary for inclusion be submitted for review and approval to a national expert panel composed of domain specialists.
- *Description of the PISA translation procedures.* It was required that national version(s) be developed by double translation and reconciliation with the source material. It was recommended that one independent translator use the English source version and that the second use the French version. In countries where the NPM had difficulty appointing competent French translators, double translation from English only was considered acceptable.

Other sections of the PISA translation guidelines were more directly intended for use by the national translators and reconciler(s):

- Recommendations to prevent common translation traps – a rather extensive section giving detailed examples on problems frequently encountered when translating assessment materials and advice on how to avoid them;
- Instructions on how to adapt the material to the national context, listing a variety of rules identifying acceptable/unacceptable national adaptations;
- Special notes on translating mathematics and science material;
- Special notes on translating questionnaires and manuals; and
- Use of national adaptation forms, to document national adaptations included in the material.

After completion of the field trial, an additional section of the guidelines was circulated to NPMs, as part of the main study manual for the NPMs, together with the revised materials to be used in the main study, to help them and their translation team with instructions on how to revise their national version(s).

TRANSLATION TRAINING SESSION

NPMs received sample materials to use for recruiting national translators and training them at the national level. The NPM meeting held in September 2001, prior to the field trial translation activities, included a workshop session for members of the national translation teams (or the person responsible for translation activities) from the participating countries. A detailed presentation was made of the material, of recommended translation procedures, of the translation guidelines and the verification process.

INTERNATIONAL VERIFICATION OF THE NATIONAL VERSIONS

As for PISA 2000, one of the most important quality control procedures implemented to ensure high quality standards in the translated assessment materials was to employ a team of independent translators, appointed and trained by the consortium, verify each national version against the English and French source versions.

Two verification co-ordination centres were established. One was at the ACER in Melbourne, for national adaptations used in the English-speaking countries. The second one was at cApStAn, a translation firm in Brussels, for all other national versions, including the national adaptations used in the French-speaking countries. Both in PISA 2000 and in PISA 2003, cApStAn had been involved in preparing the French

source versions of the PISA materials. The firm was retained because of its familiarity with the study materials, and because of its large network of international translators, many of whom had already been involved in PISA 2000 verification activities.

The consortium undertook international verifications of all national versions in languages used in schools attended by more than 5 per cent of the country's target population. For languages used in schools attended by 5 per cent or less minorities, international-level verification was deemed unnecessary since the impact on the country results would be negligible, and verification of very low frequency languages was more feasible at national level.

For a few minority languages, national versions were only developed (and verified) in the main study phase. This was considered acceptable when a national centre had arranged with another PISA country to borrow its main study national version for their minority (*e.g.* adapting the Swedish version for the Swedish schools in Finland, the Russian version for Russian schools in Latvia), and when the minority language was considered to be a dialect that differed only slightly from the main national language (*e.g.* Nynorsk in Norway).

English- or French-speaking countries or communities were permitted to submit only national adaptation forms for verification. This was considered acceptable because these countries used national versions that were identical to the source version aside from the national adaptations.

The main criteria used to recruit translators to lead the verification of the various national versions were:

- Native command of the target language;
- Professional experience as translators from either English or French into their target language;
- Sufficient command of the second source language (either English or French) to be able to use it for cross-checks in the verification of the material;
- Familiarity with the main domain assessed (in this case, mathematics); and
- Experience as teachers and/or higher education degrees in psychology, sociology or education, if possible.

As a general rule, the same verifiers were used for homolingual versions, *i.e.* the various national versions from English, French, German, Italian and Dutch-speaking countries or communities. However, the Portuguese language differs significantly from Brazil to Portugal, and the Spanish language is not the same in Spain and in Mexico, so independent native translators had to be appointed for those countries.

In a few cases, both in the field trial and the main study verification exercises, the time constraints were too tight for a single person to meet the deadlines, and additional verifiers had to be appointed and trained.

Verifier training sessions were held in Brussels, prior to the verification of both the field trial and the main study materials. Attendees received copies of the PISA information brochure, translation guidelines, the English and French source versions of the material and a verification checklist developed by the consortium. The training session:

- Presented verifiers with PISA objectives and structure;
- Familiarised them with the material to be verified;
- Discussed extensively the translation guidelines and the verification checklist;
- Arranged scheduling and dispatch logistics; and
- Reviewed security requirements.

When made available by the countries, a first bundle of translated target material was also delivered to the verifiers, and was used during the training session in a hands-on verification workshop supervised by consortium staff.

The verification procedures were improved and strengthened in a number of respects in PISA 2003, compared to PISA 2000.

- Microsoft® Word® electronic files, rather than hard copies, were used for the verification of all materials. This approach was used in order to save time and to facilitate the revision of verified material by the NPMs, who could then use the track changes facility to accept or refuse the edits proposed by the verifier.
- At the field trial phase, the verifiers completed a semi-structured questionnaire to report on the verification findings, covering issues such as overall quality of the translation and type of problems encountered. At the main study phase, they filled in a detailed item checklist, indicating whether the changes made by the consortium to each retained test item had been correctly entered in the target version, whether national revisions or adaptations were acceptable, and whether residual errors had to be corrected. These reports were used as part of the quality control materials reviewed in the data adjudication process.
- NPMs were required to have their national adaptations forms for questionnaires and manuals approved by consortium staff before submitting them for verification along with their translated questionnaires and manuals. This was a major improvement in the procedure, due to the substantive rather than linguistic nature of most of the adaptations needed in questionnaires, which often require direct negotiation between the NPM and the consortium staff in charge of data cleaning and analyses.
- A verification step was added for the coding guides, to check on the correct implementation of late changes in the scoring instructions introduced by the consortium after the NPM coding workshops.
- Use of .pdf files rather than hard copies was recommended for the submission of all final test booklets and questionnaires for final visual check of cover pages, general instruction, layout, rendering of graphics, page and item numbering, and implementation of important corrections suggested by the verifier. A report detailing all residual problems identified by the verifier was emailed within 48 hours after reception of the .pdf files to the NPMs, so that they could implement those last edits before sending their materials to print.

TRANSLATION AND VERIFICATION PROCEDURE OUTCOMES

Possible translation errors overlooked during the development of the French source version are obviously a serious concern in a procedure that recommends use of both sources as a basis for producing all other national versions. In order to check for potential systematic biases, data from the PISA 2003 main study item analyses were used to identify all items showing even minor weaknesses in the seven English-speaking countries and the five French-speaking countries that developed their national versions by just entering national adaptations in one of the source versions provided by the consortium.[2]

Out of the 267 items used in the main study:

- One hundred and three had no problems in any of the French nor English versions, and 29 had just occasional problems in one or two of the twelve countries;
- Three appeared to have weak statistics in all versions in both the English and French groups, and another ten had weaknesses in almost all of these versions. These 13 items also appeared to have major or minor flaws in at least half of the participating countries, suggesting that the content of the item (rather than translation) probably caused the poor statistics.
- No item was found that had weaknesses in all French versions, and no flaws in any of the English versions, nor the reverse. However, some imbalance was observed for the following items:
  - R102Q04A[5] and S129Q02T had no problem in the seven English versions, but flaws in three of five French versions;
  - S133Q03 had flaws in only one of five French versions, but in six of seven English versions;
  - R104Q02 had flaws in only one of five French versions, but in five of seven English versions;
  - X430Q01 had no problem in the five French versions, but flaws in four of seven English versions;
  - M598Q01, M603Q02T, R219Q01T and R220Q06 had no problem in the five French versions, but flaws in three out of seven English versions.

The remaining 13 items had mixed patterns, with two or three countries showing weaknesses in each group, but no clear trend in favour of one of the source versions.

In fact, as shown by Table 5.1, the overall per cent of weak items was very similar in both groups of countries.

Table 5.1 ■ **Mean proportion of weak items in national versions derived from the English and French source versions of the PISA 2003 main study materials**

Group of versions	Mean proportion of weak items	SD	Version	Proportion of weak items
Adapted from the English source	0.071	0.024	AUS	0.045
			CAE	0.045
			NZL	0.045
			GBR	0.078
			IRL	0.084
			SCO	0.097
			USA	0.097
Adapted from the French source	0.060	0.020	CHF	0.026
			FRA	0.058
			BEF	0.065
			CAF	0.071
			LXF	0.078

N: 154 items (the 13 items that had weaknesses in more than 50 % of the 55 national versions were omitted from this analysis).

Quality of the national versions

A total of 55 national versions of the materials were used in the PISA 2003 main study, in 33 languages. The languages were: Arabic, Bahasa Indonesian, Basque, Bokmål, Catalan, Chinese, Czech, Danish, Dutch (two national versions); English (seven versions), Finnish, French (five versions), Galician, German (four versions), Greek, Italian (two versions), Hungarian (three versions), Icelandic, Irish, Japanese, Korean, Latvian, Nynorsk, Polish, Portuguese (two versions), Serb, Slovak, Russian (two versions), Spanish (three versions), Swedish (two versions), Thai, Turkish and Valencian.

International verification was not implemented for:

- Galician and Valencian, which were used in Spain for minorities that made up less than 5 per cent of the PISA target population and which therefore were verified by national verifiers;
- Irish, which was used in Ireland for a minority that made up less than 5 per cent of PISA target population and which therefore was verified by national verifiers;
- German versions used in Belgium (German Community), Italy (Bolzano Region) and Liechtenstein, as in these cases German booklets were borrowed from Germany, Austria and Switzerland, respectively, without including any national adaptation; and
- The Chinese versions used in Macao-China, which was sourced without change from Hong Kong-China.

A few other minority versions were only borrowed, adapted and verified at the main study phase. This was the case for the Swedish version adapted by Finland for use in their Swedish schools, the Russian version adapted in Latvia, the Hungarian version adapted in Serbia and in the Slovak Republic and the Nynorsk version adapted from Bokmål in Norway.

All other versions underwent international verification both at the field trial and at the main study phase.

In their field trial reports, all NPMs were asked to describe the procedures used for the development of their national version(s). According to these reports, the procedures used for the 52 versions that underwent verification were as follows:

- Use of one of the source versions, with national adaptations: 12 national versions
- Development of a common version (through double translation from the English source and cross-check against the French source), then each country entered national adaptations: 4 national versions
- Double translation from the English and French source versions: 15 national versions
- Double translation from the English source: 15 national versions
- Double translation from the French source: 1 national version
- Use of a verified version borrowed from another PISA country, with national adaptations: 7 national versions

No case of non-standard procedure (for example, single translation) was reported, although in a small number of cases the initial versions received at the field trial phase contained so little evidence of reconciliation work that the verifier suspected the reconciler had just selected, for each test unit, one of the two independent versions received from the translators.[4]

Overall, the reports received from the verifiers confirmed that, with a few exceptions, the national versions submitted were from good to very high quality. In some of the countries that had submitted relatively poor translations in PISA 2000, significant improvements could be observed.

However, as in PISA 2000, the verification exercise proved to be an essential mechanism for ensuring quality. In virtually all versions, the verifiers identified (often many) errors that would have seriously affected the functioning of specific items – mistranslations, omissions, loan translations or awkward expressions, incorrect terminology, poor rendering of graphics or layout, errors in numerical data, grammar and spelling errors.

Two issues appeared to be of particular concern in PISA 2003, both related to the materials from PISA 2000 to be included as link items in the new study.

First, in a surprisingly large number of countries, it proved to be somewhat difficult to retrieve the electronic files containing the final national version of the materials used in the PISA 2000 main study, from which the link items had to be drawn. The verification team had to systematically check that the link units submitted were those actually used in the PISA 2000 test booklets, rather than some intermediate version where the latest edits had not been entered. In a few cases (particularly in some of the countries where the management teams for PISA 2000 and PISA 2003 were different), the verification team or the consortium had to assist by providing the correct national versions from their own central archives.

Second, it proved quite hard to prevent both the NPMs and the verifiers from correcting possible residual errors or equivalence problems identified in the link items – despite all warnings that link material should be kept identical to that used in PISA 2000. In a number of versions, at least a few minor edits were introduced, not all of which were documented. As a result, systematic queries had to be sent to the participating countries in order to check that the items to be used in the trends analyses had not undergone any unexpected change.

To prevent this type of problem in future studies, the central archive at ACER will be improved to host copies of all final national versions of the materials used in each assessment; NPMs will be asked to download the files containing their link materials from the central archive rather than retrieving them from their own directories.

Empirical evidence on the quality of the national versions obtained was collected by analysing the proportion of weak items in each national data set, based again on the PISA 2003 main study item analyses, and using the same criteria for identifying weak items as for the source versions.

The results, presented in Table 5.2, suggest that in a vast majority of the national versions, the proportion of weak items was very similar to that observed in the English and French versions directly derived from the source materials – in a number of them, the proportion of flaws even appeared to be less than in any of the English and French versions.

Of particular interest is the low proportion of weak items in the versions commonly developed by the German-speaking countries. Similar results were observed in PISA 2000 for this group of countries, which confirms that close co-operation between countries sharing a common language can be a very cost-effective way for producing high quality translations.

Table 5.2 ■ **Proportion of weak items in national versions of PISA 2003 main study materials, by translation method**

Translation/adaptation method	Mean proportion of weak items	SD	Version	Proportion of weak items
A. Adapted from the English source	0.071	0.024		
B. Adapted from the French source	0.060	0.020		
C. Adapted from the common German version	0.063	0.019	CHG	0.039
			LXG	0.058
			AUT	0.078
			GER	0.078
D. Double-translated from both English and French	0.085	0.048	BEN	0.026
			ITA	0.026
			CHI	0.052
			POL	0.052
			GRC	0.065
			DNK	0.071
			NOR	0.071
			PRT	0.071
			SWE	0.071
			ISL	0.078
			SVK	0.091
			NLD	0.104
			KOR	0.143
			BRA	0.169
			JPN	0.188
E. Double-translated from one of the source versions (English or French)	0.125	0.083	FIN	0.032
			ESC	0.039
			LVL	0.045
			ESS	0.065
			HUN	0.065
			CZE	0.071
			RUS	0.078
			HKG	0.117
			YUG	0.117
			TUR	0.123
			MAC	0.143
			MEX	0.143
			ESB	0.182
			THA	0.188
			IDN	0.266
			TUN	0.325
F. Borrowed versions	0.068	0.014	URY	0.058
			LVR	0.078

N: 154 items (the 13 items that had weaknesses in more than 50 % of the 55 national versions were omitted from this analysis).

Note: Due to the small numbers of observations for some of the minority versions (*i.e.* the Hungarian versions used in Slovakia and Serbia, the Swedish version used in Finland and the Nynorsk version used in Norway), no separate item analysis was available for these versions.

© OECD 2005

The data in Table 5.2 also seem to indicate that double-translation from only one source language (group E) may be less effective than double translation from both languages (group D), confirming a trend already observed in PISA 2000. However, the between-groups difference was not statistically significant, due to very large within-group variations in the proportion of weak items.

In fact, both group D and group E comprised a number of outliers, that is, versions that had much higher proportions of weak items than in the vast majority of other participating countries. These outliers were the national versions from the Basque Country, Brazil, Indonesia, Japan, Korea, Macao-China, Mexico, Thailand and Tunisia, as well as to a lesser extent Hong Kong-China, Serbia and Turkey.

Translation problems are not necessarily the only possible explanation for the weak item statistics in the case of outliers. It must be noted that most of these countries achieved either particularly high or particularly low average scores in all domains assessed in PISA 2003: Hong Kong-China, Korea, Japan and Macao-China consistently ranked among the best performing countries, while Brazil, Indonesia, Mexico, Thailand, Tunisia, Turkey and Serbia consistently ranked near the bottom of the distribution. In part, the greater instability of scores at both ends of the scale may perhaps explain the weaker item statistics in this particular group of countries: the PISA tests may simply have been less discriminating in those countries than in others (although other high achieving countries, such as Finland or Netherlands, and other low performers such as Uruguay and the Russian Federation, did not show the same problems in their item analyses).

However, a second possible explanation might be of some concern in terms of linguistic and cultural equivalence, *i.e.* the fact that the group of outliers included all but two of the ten PISA versions that were developed in non-Indo-European languages (Arabic, Turkish, Basque, Japanese, Korean, Thai, Chinese and Bahasa Indonesian). The two exceptions were the Finno-Ougrian languages (neither the Finnish nor the Hungarian version had poor statistics). Here again, the cultural explanation can only be partial, since some of the versions in Indo-European languages, such as the Mexican and Brazilian versions, also had high proportions of weak items.

And, finally, a third explanation may well be that competent translators and verifiers from English and French are simply harder to find in certain countries or for certain languages than for others.

Summary of items lost due to translation errors

In all cases when large DIFs or other serious flaws were identified in specific items, the NPMs were asked to review their translation of the item and to provide the consortium with possible explanations.

As often happens in this kind of exercise, no obvious translation error was found in a majority of cases. However, some residual errors could be identified, that had been overlooked by both the NPMs and the verifier. A total of 21 items were deleted from the computation of national scores for the following reasons:[5]

- Mistranslations or too confusing translations (12 items);
- Poor printing (three items);
- Failure to include some errata, or to enter them in all of the booklets containing the same unit (two items);
- Typographical errors in key numerical data (two items);

- Omission of key words (one item); and
- Failure to adequately adapt some mathematical convention (1 item).

Similarly, eight questions had to be dropped from the questionnaire data, due to omission of essential instructions from the stem (six cases), to inclusion of a negation that reversed the meaning of an item (one case), or to a confusing translation (one item).

Notes

1 Mrs. Josette Le Coq, Inspectrice, Académie de Versailles, France.

2 For the purpose of this analysis, items were considered as weak when they had one or more of the following problems: negative or positive DIF more than 0.8; discrimination less than 0.30; and fit more than 1.15. These criteria, somewhat stricter than usual, were deliberately retained in order to identify not only items that had serious flaws, but also those that simply had a slightly imperfect functioning.

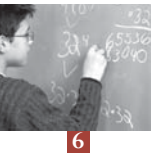
3 Item codes for reading begin with 'R', science items begin with 'S', problem-solving items begin with 'X' and mathematics items begin with 'M'.

4 One of the national versions received at the field trial phase also appeared to be a carelessly reviewed computer translation.

5 See Chapter 13.

6

Field Operations



OVERVIEW OF ROLES AND RESPONSIBILITIES

PISA was implemented in each country by a National Project Manager (NPM). The NPM implemented procedures prepared by the consortium and agreed upon by participating countries. To implement the assessment in schools, the NPMs were assisted by school co-ordinators and test administrators. Each NPM typically had several assistants, working from a base location that is referred to throughout this report as a national centre.

National Project Managers

NPMs were responsible for implementing the project within their own country. They:

- Attended NPM meetings and received training in all aspects of PISA operational procedures;
- Negotiated nationally specific aspects of the implementation of PISA with the consortium, such as national and international options, oversampling for regional comparisons, and additional analyses and reporting, *e.g.* by language group;
- Established procedures for the security of materials during all phases of the implementation;
- Prepared a series of sampling forms documenting sampling-related aspects of the national educational structure;
- Prepared the school sampling frame and submitted this to the consortium for the selection of the school sample;
- Organised for the preparation of national versions of the test instruments, questionnaires, manuals and coding guides;
- Identified school co-ordinators from each of the sampled schools and worked with them on school preparation activities;
- Selected the student sample from a list of eligible students provided by the school co-ordinators;
- Recruited and trained test administrators to administer the tests within schools;
- Nominated suitable persons to work on behalf of the consortium as external quality monitors to observe the test administration in a random selection of schools;
- Recruited and trained coders to code the open-ended items;
- Arranged for the data entry of the test and questionnaire responses, and submitted the national database of responses to the consortium; and
- Submitted a written review of PISA implementation activities following the assessment.

The PISA national project manager manual provided detailed information about the duties and responsibilities of the NPM. Supplementary manuals, with detailed information about particular aspects of the project, were also provided. These included:

- A school sampling preparation manual, which provided instructions to the NPM for documenting school sampling related issues such as the definition of the target population, school level exclusions, the proportion of small schools in the sample and so on. Instructions for the preparation of the sampling frame, *i.e.* the list of all schools containing PISA eligible students, were detailed in this manual.
- A data entry manual, which described all aspects of the use of *KeyQuest,* the data entry software prepared by the consortium, for the data entry of responses from the test booklets and questionnaires.

School co-ordinators

School co-ordinators coordinated school-related activities with the national centre and the test administrators.

The school co-ordinators:

- Established the testing date and time in consultation with the NPM;
- Prepared the student listing form with the names of all eligible students in the school and sent it to the NPM so that the NPM could select the student sample;
- Received the list of sampled students on the student tracking form from the NPM and updated it if necessary, including identifying students with disabilities or limited test language proficiency who could not take the test according to criteria established by the consortium;
- Received, distributed and collected the school questionnaire;
- Informed school staff, students and parents of the nature of the test and the test date, and secured parental permission if required by the school or education system;
- Informed the NPM and test administrator of any test date or time changes; and
- Assisted the test administrator with room arrangements for the test day.

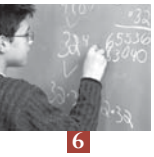
On the test day, the school co-ordinator was expected to ensure that the sampled students attended the test session(s). If necessary, the school co-ordinator also made arrangements for a follow-up session and ensured that absent students attended the follow-up session.

The consortium prepared a school co-ordinator manual that described in detail the activities and responsibilities of the school co-ordinator.

Test administrators

The test administrators were primarily responsible for administering the PISA test fairly, impartially and uniformly, in accordance with international standards and PISA procedures. To maintain fairness, a test administrator could not be the reading, mathematics or science teacher of the students being assessed, and it was preferred that they not be a staff member at any participating school. Prior to the test date, test administrators were trained by national centres. Training included a thorough review of the test administrator manual and the script to be followed during the administration of the test and questionnaire. Additional responsibilities included:

- Ensuring receipt of the testing materials from the NPM and maintaining their security;
- Co-operating fully with the school co-ordinator;
- Contacting the school co-ordinator one to two weeks prior to the test to confirm plans;
- Completing final arrangements on the test day;
- Conducting a follow-up session, if needed, in consultation with the school co-ordinator;
- Completing the student tracking form and the assessment session report form (a form designed to summarise session times, student attendance, any disturbance to the session, etc.);



- Ensuring that the number of tests and questionnaires collected from students tallied with the number sent to the school;
- Obtaining the school questionnaire from the school co-ordinator; and
- Sending the school questionnaire, the student questionnaires and all test materials (both completed and not completed) to the NPM after the testing was carried out.

The consortium prepared a test administrator manual that described in detail the activities and responsibilities of the test administrator.

THE SELECTION OF THE SCHOOL SAMPLE

The NPMs used the detailed instructions in the school sampling preparation manual to document their school sampling plan and to prepare their school sampling frame.

The national target population was defined, school and student level exclusions were identified, and aspects such as the extent of small schools and the homogeneity of students within schools were considered in the preparation of the school sampling plan.

For all but a small number of countries, the sampling frame was submitted to the consortium who selected the school sample. Having the consortium select the school sample minimised the potential for errors in the sampling process, and ensured uniformity in the outputs for more efficient data processing down the track. It also relieved the burden of this task from national centres. NPMs worked very closely with the consortium throughout the process of preparing the sampling documentation, ensuring that all nationally specific considerations related to sampling were thoroughly documented and incorporated into the school sampling plan.

While all countries were required to thoroughly document their school sampling plan, a small number of countries were permitted to select the school sample themselves. In these cases, the national centre was required to explain in detail the sampling methods used, to ensure that they were consistent with those used by the consortium. The software used was submitted to the consortium for checking. The consortium ran checks on the software and the methods as described, prior to approving the school sample selection. Further details about sampling for the main study are provided in Chapter 4.

PREPARATION OF TEST BOOKLETS, QUESTIONNAIRES AND MANUALS

As described in Chapter 2, 13 different test booklets had to be assembled with clusters of test items arranged according to the test booklet design specified by the consortium. Test items were presented in units (stimulus material and items relating to the stimulus), and each cluster contained several units. Test units and questionnaire items were initially sent to NPMs several months before the testing dates, to enable translation to begin. Units allocated to clusters, and clusters allocated to booklets, were provided a few weeks later, together with detailed instructions to NPMs about how to assemble their translated or adapted clusters into booklets.

For reference, master hard copies of all booklets were provided to NPMs, and master copies in both English and French were also available through a secure website. NPMs were encouraged to use the cover design provided by the OECD (both black and white and coloured versions of the cover design were made available). In formatting translated or adapted test booklets, NPMs had to follow as far as possible the layout in the English master copies, including allocation of items to pages. A slightly smaller

or larger font than in the master copy was permitted if it was necessary to ensure the same page layout as that of the source version.

NPMs were required to submit their cognitive material in unit form, along with a form documenting any proposed national adaptations, for verification by the consortium. NPMs incorporated feedback from the verifier into their material and assembled the test booklets. These were submitted once more to the consortium, who performed a final visual check of the materials. This was a verification of the layout, instructions to the student, the rendering of graphic material, etc. Once feedback from the final optical check had been received and incorporated into the test booklets, the NPM was ready to send the materials to print.

The student questionnaire contained one, two, or three modules, according to whether the international options of Information Commumnication Technology (ICT) and Educational Career questionnnaires were being added to the core component. About half the countries chose to administer the Educational Career component and just over three-quarters used the ICT component. The core component had to be presented first in the questionnaire booklet. If both international options were used, the Educational Career module was to be placed ahead of the ICT module.

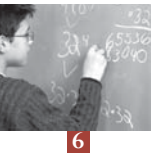
As with the test material, source versions of the questionnaire instruments in both English and French were provided to NPMs to be used to assist in the translation of this material.

NPMs were permitted to add questions of national interest as national options to the questionnaires. Proposals and text for these were submitted to the consortium for approval as part of the process of reviewing adaptations to the questionnaires. It was recommended that the additional material should be placed at the end of the international modules. The student questionnaire was modified more often than the school questionnaire.

NPMs were required to submit a form documenting all proposed national adaptations to questionnaire items to the consortium for approval. Following approval of adaptations, the consortium verified the material. NPMs implemented feedback from verification in the assembly of their questionnaires. The questionnaires were then submitted once more in order to conduct a final optical check. Following feedback from the final optical check, NPMs made final changes to their questionnaires prior to printing.

The school co-ordinator and test administrator manuals also had to be translated into the national languages so English and French source versions of each manual were provided by the consortium. NPMs were required to submit a form documenting all proposed national adaptations to the manuals for the consortium's approval. Following approval of the adaptations, the manuals were prepared and submitted to the consortium for verification. NPMs implemented feedback from the verifier into their manuals prior to printing. A final optical check was not required for the manuals.

In countries with multiple languages, the test instruments and manuals had to be translated into each test language. In a small number of bilingual countries, where test administrators were familiar with both test languages, only one national version of their manual was developed. However in these cases it was a requirement that the test script, included within the test administrator manual, was translated into both languages of the test.



THE SELECTION OF THE WITHIN-SCHOOL SAMPLE

Following the selection of the school sample by the consortium, the list of sampled schools was returned to national centres. NPMs then contacted these schools and requested a list of all PISA-eligible students from each school. This was provided on the student listing form, and was used by NPMs to select the within-school sample.

NPMs were required in most cases to select the student sample using *KeyQuest*, the PISA student sampling and data entry software prepared by the consortium (see Chapter 4). *KeyQuest* generated the list of sampled students for each school, known as the student tracking form that served as the central administration document for the study and linked students, test booklets and student questionnaires.

Only in exceptional circumstances were NPMs permitted to select their student sample without using *KeyQuest*. In these cases, NPMs were required to detail the sampling methods that were used, and the consortium verified these.

PACKAGING AND SHIPPING MATERIALS TO THE SCHOOLS

NPMs were allowed some flexibility in how the materials were packaged and distributed, depending on national circumstances. Regardless of how materials were packaged and shipped, the following were sent either to the test administrator or to the school:

- Test booklets and student questionnaires for the number of students sampled;
- The student tracking form;
- Two copies of the assessment session report form;
- The packing form;
- The return shipment form;
- Additional materials, such as rulers and calculators; and
- Additional school and student questionnaires, as well as a number of extra test booklets.

Of the 13 separate test booklets, one was pre-allocated to each student by the *KeyQuest* software from a random starting point in each school. *KeyQuest* was then used to generate the school's student tracking form, which contained the number of the allocated booklet alongside each sampled student's name.

It was recommended that labels be printed, each with a student identification number and the test booklet number allocated to that student. If it was as an acceptable procedure within the country then it was also recommended that the student's name be printed on the label. Two or three copies of each student's label could be printed, and used to identify the test booklet, the questionnaire and a packing envelope if used.

It was specified that the test booklets for a school be packaged so that they remained secure, possibly by wrapping them in clear plastic and then heat-sealing the package, or by sealing each booklet in a labelled envelope. Three scenarios were illustrative of acceptable approaches to the packaging and shipping the assessment materials:

© OECD 2005

- *Country A:* all assessment materials shipped directly to the schools; school staff (not teachers of the students in the assessment) to conduct the testing sessions; materials assigned to students before packaging; materials labelled and then sealed in envelopes also labelled with the students' names and identification numbers.
- *Country B:* materials shipped directly to the schools; external test administrators employed by the national centre to administer the tests; the order of the booklets in each bundle matches the order on the student tracking form; after the assessment has been completed, booklets are inserted into envelopes labelled with the students' names and identification numbers and sealed.
- *Country C:* materials shipped to test administrators employed by the national centre; bundles of 35 booklets sealed in plastic, so that the number of booklets can be checked without opening the packages; test administrators open the bundle immediately prior to the session and label the booklets with the students' names and ID numbers from the student tracking form.

RECEIPT OF MATERIALS AT THE NATIONAL CENTRE AFTER TESTING

The consortium recommended that the national centre establish a database of schools before testing began to record the shipment of materials to and from schools, to keep track of materials sent and returned, and to monitor the progress of the materials throughout the various steps in processing booklets after the testing.

The consortium recommended that upon receipt of materials back from schools, the counts of completed and unused booklets also be checked against the participation status information recorded on the student tracking form by the test administrator.

CODING OF THE TESTS AND QUESTIONNAIRES

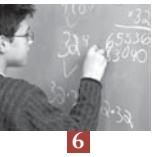
This section describes PISA's coding procedures, including multiple coding (as required for inter-coder reliability studies), and also makes brief reference to pre-coding of responses to a few items in the student questionnaire. For each domain, the proportions of student responses requiring manual evaluation were as follows: mathematics, 42 per cent; reading, 61 per cent; science, 43 per cent; and problem solving, 58 per cent. Overall, 47 per cent of items across all domains (mathematics, reading, science and problem solving) required manual coding by trained coders.

This was a complex operation, as booklets had to be randomly assigned to coders and, for the minimum recommended sample size per country of 4 500 students, more than 140 000 responses had to be evaluated.

It is crucial for comparability of results in a study such as PISA that students' responses are scored uniformly from coder to coder, and from country to country. Comprehensive criteria for coding, including many examples of acceptable and not acceptable responses, for each of reading, mathematics, science and problem solving, were prepared by the consortium and provided to NPMs in the coding guides.

Preparing for coding

In setting up the coding of students' responses to open-ended items, NPMs had to carry out or oversee several steps:



- The adaption or translation of the coding guides as needed and their submission to the consortium for verification;
- The recruitment and training of coders;
- The location of suitable local examples of responses to use in training;
- The organisation of booklets as they were returned from schools;
- The selection of booklets for multiple coding;
- The overseeing of the coding the booklets according to the international design;
- The assignment of multiple codes to a selected sub-sample of booklets once the single coding was completed; and
- The submission of a sub-sample of booklets for the inter-country coder reliability study (see Chapter 10).

Detailed instructions for each step were provided in the national project manager manual. Key aspects of the process are included here.

International training

Representatives from each national centre were required to attend two international coder training sessions – one immediately prior to the field trial and one immediately prior to the main study. At the training sessions, consortium staff familiarised national centre staff with the coding guides and their interpretation.

Staffing

NPMs were responsible for recruiting appropriately qualified people to carry out the single and multiple coding of the test booklets. In some countries, pools of experienced coders from other projects could be called on. It was not necessary for coders to have high-level academic qualifications, but they needed to have a good understanding of either mid-secondary level mathematics and science or the language of the test, and to be familiar with ways in which secondary-level students express themselves. Teachers on leave, recently retired teachers and senior teacher trainees were all considered to be potentially suitable coders. An important factor in recruiting coders was that they could commit their time to the project for the duration of the coding, which was expected to take up to two months.

The consortium provided a coder recruitment kit to assist NPMs in screening applicants. These materials were similar in nature to the coding guides, but were much briefer. They were designed so that potential applicants could be given a brief training session, after which they coded some student responses. Guidelines for assessing the results of this exercise were supplied. The materials also provided applicants with the opportunity to assess their own suitability for the task. The number of coders required was governed by the design for multiple coding (described in a later section). For the main study, it was recommended to have 16 coders across the domains of mathematics, science and problem solving, and an additional eight coders for reading. These numbers of coders were considered to be adequate to meet the timeline, of submitting their data within three months of testing, for countries testing between 4 500 (the minimum number required) and 6 000 students.

For larger numbers of students, or in cases where coders would code across different combinations of domains, NPMs could prepare their own design and submit it to the consortium for approval. A minimum of four coders were required in each domain to satisfy the requirements of the multiple coding design.

Given that several weeks were required to complete the coding, it was recommended that back-up coders (at least two for mathematics, science and problem solving, and one for reading) be trained and included in at least some of the coding sessions.

The coding process was complex enough to require a full-time overall supervisor who was familiar with the logistical aspects of the coding design, the procedures for checking coder reliability, the coding schedules and also the content of the tests and the coding guides.

NPMs were also required to designate persons with subject-matter expertise, familiarity with the PISA tests and, if possible, experience in coding student responses to open-ended items, in order to act as table leaders during the coding. Table leaders were expected to participate in the actual coding and spend extra time monitoring consistency. Good table leaders were essential to the quality of the coding, as their main role was to monitor coders' consistency in applying the coding criteria. They also assisted with the flow of booklets, and fielded and resolved queries about the coding guides and about particular student responses in relation to the guides, consulting the supervisor as necessary when queries could not be resolved. The supervisor was then responsible for checking such queries with the consortium.

Several persons were needed to unpack, check and assemble booklets into labelled bundles so that coders could respect the specified design for randomly allocating sets of booklets to coders.

Consortium coding query service

A coding query service was provided by the consortium to clarify any questions about particular items that could not be resolved at the national centre level. Responses to coding queries were placed on the secure website, which was accessible to the NPMs from all participating countries. There were 376 queries in mathematics, 345 in reading, 135 in science and 61 in problem solving.

Confidentiality forms

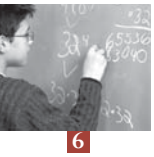
Before seeing or receiving any copies of PISA test materials, prospective coders were required to sign a confidentiality form, obligating them not to disclose the content of the PISA tests beyond the groups of coders and trainers with whom they would be working.

National training

Anyone who coded the PISA main survey test booklets had to participate in specific training sessions, regardless of whether they had had related experience or had been involved in the PISA field trial coding. To assist NPMs in carrying out the training, the consortium prepared training materials in addition to the detailed coding guides. Training within a country could be carried out by the NPM or by one or more knowledgeable persons appointed by the NPM. Subject matter knowledge was important for the trainer, as was an understanding of the procedures, which usually meant that more than one person was involved in leading the training.

The recommended allocation of booklets to coders assumed coding by cluster. This involved completing the coding of each item separately within a cluster within all of the booklets allocated to the coder, before moving to the next item and completing one cluster before moving to the next.

Coders were trained by cluster for the seven mathematics clusters and two clusters of reading, science and problem solving. During a training session, the trainer reviewed the coding guide for a cluster of units with the coders, then had the coders assign codes to some sample items for which the appropriate codes had been



supplied by the consortium. The trainer reviewed the results with the group, allowing time for discussion, querying and clarification of reasons for the pre-assigned codes. Trainees then proceeded to code independently some local examples that had been carefully selected by the trainer in conjunction with national centre staff. It was recommended that prospective coders be informed at the beginning of training that they would be expected to apply the coding guides with a high level of consistency, and that reliability checks would be made frequently by table leaders and the overall supervisor as part of the coding process.

Ideally, table leaders were trained before the larger groups of coders since they needed to be thoroughly familiar with both the test items and the coding guides. The coding supervisor explained these to the point where the table leaders could code and reach a consensus on the selected local examples to be used later with the larger group of trainees. They also participated in the training sessions with the rest of the coders, partly to strengthen their own knowledge of the coding guides and partly to assist the supervisor in discussions with the trainees of their pre-agreed codes to the sample items. Table leaders received additional training in the procedures for monitoring the consistency with which coders applied the criteria.

Length of coding sessions

Coding responses to open-ended items is mentally demanding, requiring a level of concentration that cannot be maintained for long periods of time. It was therefore recommended that coders work for no more than six hours per day on actual coding and that they take two or three breaks for coffee and lunch. Table leaders needed to work longer on most days so that they had adequate time for their monitoring activities.

Logistics prior to coding

Sorting booklets

When booklets arrived back at the national centre, they were first tallied and checked against the session participation codes on the student tracking form. Used and unused booklets were separated; used booklets were sorted by student identification number (if they had not been sent back in that order) and were then separated by booklet number. School bundles were kept in school identification order, filling in sequence gaps as packages arrived. student tracking forms were carefully filed in ring binders in school identification order. If the school identification number order did not correspond with the alphabetical order of school names, it was recommended that an index of school name against school identification be prepared and kept with the binders.

Because of the time frame for countries to have all their coding done and data submitted to the consortium, it was usually impossible to wait for all materials to reach the national centre before beginning to code. In order to manage the design for allocating booklets to coders, however, it was recommended to start coding only when at least half of the booklets had been returned.

Selection of booklets for multiple coding

Each country was required to set aside 100 of each of booklets 1 to 6, 8, 10 and 12 for multiple coding. The first two clusters from each of these booklets were multiple coded. This arrangement ensured that all clusters were included in the multiple coding.

The main principle in setting aside the booklets for multiple coding was that the selection needed to ensure a wide spread of schools and students across the whole sample and to be random as far as possible. The

simplest method for carrying out the selection was to use a ratio approach based on the expected total number of completed booklets.

With a sample of around 5 000 students, approximately 400 of each booklet type would have been returned from schools. Therefore, approximately one in four of each of the booklet types used in multiple coding needed to be selected for the multiple coding activity. With a larger sample size, the selection ratios needed to be adjusted so that the correct numbers of each booklet were selected from the full range of participating schools.

In a country where booklets were provided in more than one language, if the language represented 20% or more of the target population, the 900 booklets to be set aside for multiple coding were allocated in proportion to the language group. Multiple coding was not required for languages representing less than 20% of the target population.

Booklets for single coding

Single coding was required for the booklets remaining after those for multiple coding had been set aside, as well as for the third and fourth clusters from those set aside for multiple coding. Some items requiring coding did not need to be included in the multiple coding. These were closed-constructed response items that required a coder to assign a right or wrong code, but did not require any coder judgement. The last coder in the multiple-coding process coded these items in the booklets set aside for multiple coding, as well as the items requiring single coding from the third and fourth clusters. Other items such as multiple-choice response items required no coding and were directly data-entered.

How codes were shown

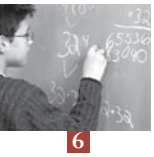
A string of small code numbers corresponding to the possible codes for the item as delineated in the relevant coding guide appeared in the upper right-hand side of each item in the test booklets that required coding. For booklets being processed by a single coder, the code assigned was indicated directly in the booklet by circling the appropriate code number alongside the item. Tailored coding record sheets were prepared for each booklet for the multiple coding and used by all but the last coder so that each coder undertaking multiple coding did not know which codes other coders had assigned.

For the reading clusters, item codes were often just 0, 1 and 9, indicating incorrect, correct and missing, respectively. Provision was made for some of the open-ended items to be coded as partially correct, usually with '2' as fully correct and '1' as partially correct, but occasionally with three degrees of correctness indicated by codes of '1', '2' and '3'.

For the mathematics, problem-solving and science clusters, a two-digit coding scheme was adopted for the items requiring constructed-responses. The first digit represented the 'degree of correctness' code, as in reading; the second indicated the content of the response or the type of solution method used by the student. Two-digit codes were originally proposed by Norway for TIMSS and were adopted in PISA because of their potential for use in studies of student learning and thinking.

Coder identification numbers

Coder identification numbers were assigned according to a standard three-digit format specified by the consortium. The first digit showed the combination of domains that the coder would be working across, and the second and third digits had to uniquely identify the coders within their set. For example,



16 coders coding across the three domains of mathematics, science and problem solving were given identification numbers 701 to 716. Eight coders who coded just reading were given identification numbers 201 to 208. Coder identification numbers were used for two purposes: implementing the design for allocating booklets to coders, and in monitoring coder consistency in the multiple-coding exercises.

Single coding design

Single coding of mathematics, science and problem solving

In order to code by cluster, each coder needed to handle 3 or 4 of the 13 booklet types at a time, depending on the number of booklet types that the cluster appeared in. For example, mathematics cluster 1 occurred in booklets 1, 5, 11 and 13. Each of these appearances had to be coded before another cluster was started. Moreover, since coding was done item by item, the item was coded across these different booklet types before the next item was coded.

A design to ensure the random allocation of booklets to coders was prepared based on the recommended number of 16 coders and the minimum sample size of 4 500 students from 150 schools. Booklets were to be sorted by student identification within schools. With 150 schools and 16 coders, each coder had to code a cluster within a booklet from eight or nine schools ($150/16 \approx 9$). Table 6.1 shows how booklets needed to be assigned to coders for the single coding. Further explanation of the information in this table is presented below.

Table 6.1 ■ **Design for the single coding of mathematics, science and problem solving**

Cluster	Booklets	School subsets															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
M1	1,5,11,13	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716
M2	1,2,6,12	716	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715
M3	2,3,7,13	715	716	701	702	703	704	705	706	707	708	709	710	711	712	713	714
M4	1,3,4,8	714	715	716	701	702	703	704	705	706	707	708	709	710	711	712	713
M5	2,4,5,9	713	714	715	716	701	702	703	704	705	706	707	708	709	710	711	712
M6	3,5,6,10	712	713	714	715	716	701	702	703	704	705	706	707	708	709	710	711
M7	4,6,7,11	711	712	713	714	715	716	701	702	703	704	705	706	707	708	709	710
S1	5,7,8,12	710	711	712	713	714	715	716	701	702	703	704	705	706	707	708	709
S2	6,8,9,13	709	710	711	712	713	714	715	716	701	702	703	704	705	706	707	708
PS1	3,9,11,12	708	709	710	711	712	713	714	715	716	701	702	703	704	705	706	707
PS2	4,10,12,13	707	708	709	710	711	712	713	714	715	716	701	702	703	704	705	706

According to this design, cluster M1 in subset 1 (schools 1 to 9) was to be coded by coder 701. Cluster M1 in subset 2 (schools 10 to 18) was to be coded by coder 702, and so on. For cluster M2, coder 701 was to code all booklets from subset 2 (schools 10 to 18), coder 702 was to code all booklets from subset 3 (schools 19 to 27) and so on. Subset 1 of cluster M2 (schools 1 to 9) was to be coded by coder 716.

If booklets from all participating schools were available before the coding began, the following steps would be involved in implementing the design:

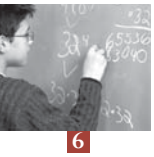
- *Step 1*: Set aside booklets for multiple coding and then divide the remaining booklets into school subsets as above; (subset 1: schools 1 to 9; subset 2: schools 10 to 18, etc., to achieve 16 subsets of schools).
- *Step 2*: Assuming that coding begins with cluster M1: coder 701 takes booklets 1, 5, 11 and 13 for school subset 1; coder 702 takes booklets 1, 5, 11 and 13 for school subset 2; etc.; until coder 716 takes booklets 1, 5, 11 and 13 for school subset 16.
- *Step 3*: Coders code all of the first cluster M1 item requiring coding in the booklets that they have.
- *Step 4*: The second cluster M1 item is coded in all four booklet types, followed by the third cluster M1 item, etc., until all cluster M1 items are coded.
- *Step 5*: For cluster M2, as per the row of the table in Table 6.1 corresponding to M2 in the left-most column, each coder is allocated a subset of schools different from their subset for cluster M1. Coding proceeds item by item within the cluster.
- *Step 6*: For the remaining clusters, the rows corresponding to M3, M4, etc., in the table are followed in succession.

Single coding of reading

Table 6.2 shows a similar design that was prepared for the single coding of reading. As the recommended number of coders for reading (8) was half that recommended for coding mathematics, science and problem solving, each coder was allocated 'two subsets worth' of schools. Also, as there were just two different clusters of reading, each of which appeared in four booklet types, each coder coded two of the four appearances of a cluster. This ensured that a wider range of coders was used for each school subset. For the coding of cluster R1, for example, coder 201 coded the appearances of this cluster in booklets 1 and 7 from school subsets 1 and 2 (*i.e.* schools 1-18), coder 202 coded this cluster from booklets 1 and 7 for school subsets 3 and 4, and so on. For the other two appearances of cluster R1 (booklets 9 and 10), coder 203 coded these from school subsets 1 and 2, coder 204 from school subsets 3 and 4, and so on.

Table 6.2 ■ **Design for the single coding of reading**

Cluster	Booklets	School subsets							
		1 - 2	3 - 4	5 - 6	7 - 8	9 -10	11 - 12	13 - 14	15 - 16
R1	1,7	201	202	203	204	205	206	207	208
R1	9,10	203	204	205	206	207	208	201	202
R2	2,8	205	206	207	208	201	202	203	204
R2	10,11	207	208	201	202	203	204	205	206



As a result of this procedure, the booklets from each subset of schools were processed by 15 different coders, one for each distinct cluster of mathematics, problem solving and science, and two for each cluster of reading. Also each student's booklet was coded by four different coders, one for each of the four clusters in the student's booklet. Spreading booklets among coders in this way minimised the effects of any systematic leniency or harshness in coding.

In practice, most countries would not have had completed test booklets back from all their sampled schools before coding needed to begin. NPMs were encouraged to organise the coding in two waves, so that it could begin after materials were received back from one-half of their schools. Schools would not have been able to be assigned to school sets for coding exactly in their school identification order, but rather by identification order combined with when their materials were received and processed at the national centre.

Une Heure (UH) booklet

Countries using the shorter, special purpose UH booklet were advised to process this separately from the remaining booklets. Small numbers of students used this booklet, only a few items required coding, and they were not arranged in clusters. NPMs were cautioned that booklets needed to be allocated to several coders to ensure uniform application of the coding criteria for UH booklet, as for the main coding.

Multiple coding

For PISA 2003, four coders independently coded all short response and open constructed-response items from the first half of a sample of booklets. 100 of each of booklets 1 to 6, 8, 10 and 12, a total of 900 booklets were selected for this multiple coding activity. Multiple coding was done at or towards the end of the coding period, after coders had familiarised themselves with and were confident in using the coding guides. As noted earlier, the first three coders of the selected booklets circled codes on separate record sheets, tailored to booklet type and domain (reading, mathematics, science or problem solving), using one page per student. The coding supervisor checked that coders correctly entered student identification numbers and their own identification number on the sheets; this was crucial to data quality. The UH booklet was not included in the multiple coding.

While coders would have been thoroughly familiar with the coding guides by the time of multiple coding, they may have most recently coded a different booklet from those allocated to them for multiple coding. For this reason, they needed to have time to re-read the relevant coding guide before beginning the coding. It was recommended that time be allocated for coders to refresh their familiarity with the guides and to look again at the additional practice material before proceeding with the multiple coding. As in the single coding, coding was to be done item by item. For manageability, items from the four clusters within a booklet type (*e.g.* booklet 1) were coded before moving to another booklet type, rather than coding by cluster across several booklet types. It was considered that coders would be experienced enough in applying the coding criteria by this time that coding by booklet would be unlikely to detract from the quality of the data.

Multiple coding of mathematics, science and problem solving

The specified multiple coding design for mathematics, science and problem solving, shown in Table 6.3, assumed 16 coders with identification numbers 701 to 716. The importance of following the design exactly as specified was stressed, as it provided for balanced links between clusters and

coders. Table 6.3 shows 16 coders grouped into four groups of four, with Group 1 comprising the first four coders (701-704), Group 2 the next four (705-708), etc. The design involved two steps, with the booklets divided into two sets. booklets 1 to 4 made up one set, and booklets 5, 6, 8 and 12 the second set. The four codings were to be carried out by rotating the booklets to the four coders within each group.

Table 6.3 ■ **Design for the multiple coding of mathematics, science and problem solving**

Step	Booklet	Coder IDs	Clusters for multiple coding	Clusters for single coding
1	1	701, 702, 703, 704	M1,M2	M4
	2	705, 706, 707, 708	M2,M3	M5
	3	709, 710, 711, 712	M3,M4	M6,PS1
	4	713, 714, 715, 716	M4,M5	M7,PS2
2	5	703, 704, 705, 706	M5, M6	S1, M1
	6	707, 708, 709, 710	M6, M7	S2, M2
	8	711, 712, 713, 714	S1, S2	M4
	12	715, 716, 701, 702	PS1, PS2	M2, S1
3	10	Unspecified		PS2, M6

In this scenario, with all 16 coders working, booklets 1 to 4 were to be coded at the same time in the first step. The 100 copies of booklet 1, for example, were to be divided into four bundles of 25, and rotated among coders 701, 702, 703 and 704, so that each coder eventually would have coded clusters M1 and M2 from all of the 100 booklets. At the fourth rotation, after each coder had finished the multiple coding of clusters M1 and M2 from the 25 booklets in their pile, they would then single code any mathematics, science or problem-solving clusters from the second half of the booklet. The same pattern was to be followed for booklets 2, 3 and 4.

After booklets 1 to 4 had been put through the multiple-coding process, the groups of coders were altered to follow the allocation shown in Step 2 of Table 6.3. That is, coders 703, 704, 705 and 706 were to code booklet 5, coders 707, 708, 709 and 710 were to code booklets 6, and so on for the remaining booklets. If only eight coders were available, the design could be applied by using the group designations in Table 6.4, however four steps, not two, with two booklets coded per step rather than four, were then needed to complete the exercise.

Allocating booklets to coders for multiple coding was quite complex and the coding supervisor had to monitor the flow of booklets throughout the process.

Multiple coding of reading

The multiple-coding design for reading shown in Table 6.4 assumed four coders, with identification numbers 201 to 204.

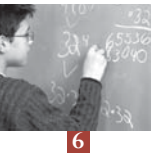


Table 6.4 ■ **Design for the multiple coding of reading**

Step	Booklet	Coder IDs	Clusters for multiple coding	Clusters for single coding
1	10	201,202,203,204	R1, R2	
2	1	Unspecified		R1
	2	Unspecified		R2
	8	Unspecified		R2

If different coders were used for mathematics science and problem solving, a different multiple-coding design was necessary. The minimum allowable number of coders coding a domain was four, in which case each booklet had to be coded by each coder.

Managing the actual coding

Booklet flow

To facilitate the flow of booklets, it was important to have ample table surfaces on which to place and arrange them by type and school subset. The bundles needed to be clearly labelled. For this purpose, it was recommended that each bundle of booklets be identified by a batch header for each booklet type (booklets 1 to 13), with spaces for the number of booklets and school identifications represented in the bundle to be written in. In addition, each header sheet was to be pre-printed with a list of the clusters in the booklet, with columns where the date and time, coder's name and identification, and table leader's initials could be entered as the bundle was coded and checked.

Separating the coding of reading and mathematics/ science/ problem solving

It was recommended that the coding of reading and the coding of mathematics, science and problem solving be done at least partly at different times (for example, reading coding could start a week or two ahead). This was to minimise the risk of different coders requiring the same booklets, so that an efficient flow of booklets through the coding process could be maintained.

Familiarising coders with the coding design

The relevant design for allocating booklets to coders was explained either during the coder training session or at the beginning of the first coding session (or both). The coding supervisor was responsible for ensuring that coders adhered to the design, and used clerical assistants if needed. Coders could better understand the process if each was provided with a card or sheet indicating the bundles of booklets to be taken and in which order.

Consulting table leaders

During the initial training, practice, and review, it was expected that coding issues would be discussed openly until coders understood the rationale for the coding criteria (or reached consensus where the coding guide was incomplete). Coders were not permitted to consult other coders or table leaders during the additional practice exercises undertaken following the training to gauge whether all or some coders needed more training and practice (see next subsection).

Following the training, coders were advised to work quietly, referring queries to their table leader rather than to their neighbours. If a particular query arose often, the table leader was advised to discuss it with the rest of the group.

For the multiple coding, coders were required to work independently without consulting other coders.

Monitoring single coding

The steps described here represented the minimum level of monitoring activities required. Countries wishing to implement more extensive monitoring procedures during single coding were encouraged to do so.

The supervisor, assisted by table leaders, was advised to collect coders' practice papers after each cluster practice session and to tabulate the codes assigned. These were then to be compared with the pre-agreed codes: each matching code was considered a hit and each discrepant code was considered a miss. To reflect an adequate standard of reliability, the ratio of hits to the total of hits plus misses needed to be 0.85 or more. In mathematics, science and problem solving, this reliability was to be assessed on the first digit of the two-digit codes. A ratio of less than 0.85, especially if lower than 0.80, was to be taken as indicating that more practice was needed, and possibly also more training.

Table leaders played a key role during each coding session and at the end of each day, by spot-checking a sample of booklets or items that had already been coded to identify problems for discussion with individual coders or with the wider group, as appropriate. All booklets that had not been set aside for multiple coding were candidates for this spot-checking. It was recommended that, if there were indications from the practice sessions that one or more particular coders might be experiencing problems in using the coding guide consistently, then more of those coders' booklets should be included in the checking. Table leaders were advised to review the results of the spot-checking with the coders at the beginning of the next day's coding. This was regarded primarily as a mentoring activity, but NPMs were advised to keep in contact with table leaders and the coding supervisor if there were individual coders who did not meet criteria of adequate reliability and would need to be removed from the pool.

Table leaders were to initial and date the header sheet of each batch of booklets for which they had carried out spot-checking. Some items and booklets from each batch and each coder had to be checked.

Cross-national coding

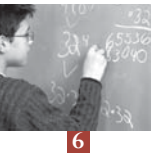
Cross-national comparability in assigning codes was explored through an inter-country coder reliability study (see Chapter 10 and Chapter 14).

Questionnaire coding

The main coding required for the student questionnaire internationally was the mother's and father's occupation and student's occupational expectation. Four-digit ISCO codes (ILO, 1990) were assigned to these three variables. In several countries, this could be done in many ways. NPMs could use a national coding scheme with more than 100 occupational title categories, provided that this national classification could be recoded to ISCO. A national classification was preferred because relationships between occupational status and achievement could then be compared within a country using both international and national measures of occupational status.

The PISA website gave a short, clear summary of ISCO codes and occupational titles for countries to translate if they had neither a national occupational classification scheme nor access to a full translation of ISCO.

In their national options, countries may also have needed to pre-code responses to some items before data from the questionnaire were entered into the software.



DATA ENTRY, DATA CHECKING AND FILE SUBMISSION

Data entry

The consortium provided participating countries with data entry software (*KeyQuest*) that ran under Microsoft® Windows 95® or later, and Microsoft® Windows NT 4.0® or later. *KeyQuest* contained the database structures for all of the booklets, questionnaires and tracking forms used in the main survey. Variables could be added or deleted as needed for national options. Approved adaptations to response categories could also be accommodated. Student response data were entered directly from the test booklets and questionnaires, except for the multiple-coding study, where the codes from the first three coders had been written on separate sheets. Information regarding the participation of students, recorded by the school co-ordinator and test administrator on the student tracking form, was entered directly into *KeyQuest*. Several questions from the session report form, such as the timing of the session, were also entered into *KeyQuest*.

KeyQuest performed validation checks as data were entered. Importing facilities were also available if data had already been entered into text files, but it was strongly recommended that data be entered directly into *KeyQuest* to take advantage of its many PISA-specific features. A *KeyQuest* manual provided generic technical details of the functionality of the *KeyQuest* software. A separate data entry manual provided complete instructions, specific to the PISA 2003 main study, regarding data entry, data management and how to carry out validity checks.

Data checking

NPMs were responsible for ensuring that many checks of the quality of their country's data were made before the data files were submitted to the consortium. These checks were explained in detail in the data entry manual, and could be simply applied using the *KeyQuest* software. The checking procedures required that the list of sampled schools and the student tracking form for each school were already accurately completed and entered into *KeyQuest*. Any errors had to be corrected before the data were submitted. Copies of the cleaning reports were to be submitted together with the data files. More details on the cleaning steps are provided in Chapter 11.

Data submission

Files to be submitted included:

- Data for the test booklets and context questionnaires;
- Data for the international option instrument(s) if used;
- Data for the multiple-coding study;
- Session report data;
- Data cleaning reports;
- The list of sampled schools; and
- Student tracking forms.

Copies, either hard or electronic, of the last two items were also required.

After data were submitted

NPMs were required to designate a data manager who would work actively with the consortium's data processing centre at ACER during the international data cleaning process. Responses to requests for information by the processing centre were required within three working days of the request.

THE MAIN STUDY REVIEW

NPMs were required to complete a structured review of their main study operations. The review was an opportunity to provide feedback to the consortium on the various aspects of the implementation of PISA, and to provide suggestions for areas that could be improved. It also provided an opportunity for the NPM to formally document aspects such as the operational structure of the national centre, the security measures that were implemented, the use of contractors for particular activities and so on.

The main study review was submitted to the consortium four weeks after the submission of the national database.

7

Monitoring the Quality of PISA

It is essential that users of the PISA data have confidence that the data collection activities have been undertaken to a high standard. The quality assurance that provides this confidence consists of two methods. The first is to carefully develop and document procedures that will result in data of the desired quality, the second is to monitor and record the implementation of the documented procedures. Should it happen that the documented processes are not fully implemented, it is necessary to understand to what extent they were not, and the likely implications for the data.

Quality monitoring is, therefore, the process of systematically observing and recording the extent to which data are collected, retrieved, and stored according to the procedures described in the field operations manuals. Quality monitoring is a continuous process that identifies potential issues and allows forestalment of operational problems. The responsibility for quality control resides with the National Project Managers (NPMs) while quality monitoring is a collaborative process between the NPM and the consortium that assists the NPM.

A comprehensive program of continuous quality monitoring was central to ensuring full, valid implementation of the PISA 2003 procedures and the recording of deviation from those procedures. The main elements of the quality monitoring procedures were:

- *Consortium experts* – To assist NPMs in the planning and implementation of key processes, consortium experts systematically monitored the key processes of school and student sampling, translation and preparation of instruments, coding of responses, field operations, and data preparation.
- *National centre quality monitors* (NCQMs) – To observe the implementation of PISA field operations at the national level, consortium representatives visited NPMs in each country.
- *PISA quality monitors* (PQMs) – Employed by the consortium and located in participating countries, PQMs visited a sample of schools to record the implementation of the documented field operations in the main study. They typically visited 15 schools in each country.
- *NPM quality surveys* – The consortium developed a series of instruments through which NPMs systematically self-reported on the implementation of key processes at the national level.
- *PISA test administration reports* – PISA test administrators completed a report after each PISA test administration, thus providing an overview of the test administration at the national level.

PREPARATION OF QUALITY MONITORING INSTRUMENTS

The purpose of quality monitoring is to observe and record the implementation of the described procedures; therefore, the field operations manuals provided the foundation for all the quality-monitoring procedures. The manuals that formed the basis for the quality-monitoring procedures were the national project manager manual, test administrator manual, school co-ordinator manual, school sampling preparation manual and the PISA data management manual. The quality monitoring instruments developed from these manuals include a range of sampling forms, a translation and verification schedule for instruments, a NCQM interview schedule, PQM instruments, NPM quality surveys, and a PISA test administrator test session report.

Sampling forms

The consortium developed a series of forms for monitoring school and student sampling activities. The NPM and consortium experts negotiated agreement on sampling plans and outcomes (see Chapter 4).

Translation and verification schedule

This is an instrument detailing the quality monitoring activities for the preparation and translation of instruments monitored instrument preparation at the national level (see Chapter 5).

National centre quality monitor interview schedule

A standard schedule was prepared by the consortium to systematically record the outcomes of the NCQM site visit. The interview schedule recorded information on:

- The general organisation of PISA in that country;
- The quality of test administrators;
- The adequacy of security and confidentiality provisions;
- The selection of the school sample;
- The selection of the student sample;
- The quality of the student tracking procedures;
- The quality of translation procedures;
- The quality of assessment booklet assembly procedures;
- The adequacy and quality of the coding procedures; and
- The independence of the PQMs.

PISA quality monitor instruments

A PQM data collection sheet was developed for PQMs to systematically record their observations during school visits. The data collection sheet recorded information on:

- The use of test script;
- The test session timing;
- The security of materials;
- The environment of the test session;
- The implementation of the student tracking procedures;
- The conduct of the students; and
- The views of the school co-ordinator.

A general observation sheet recorded their general impressions of the implementation of PISA at the national level. The general observation sheet recorded information on:

- The security of materials;
- The overall contribution of test administrators;
- The overall contribution of school co-ordinators;
- The attitude and response of students to the cognitive sessions;

- The attitude and response of students to the questionnaire session; and
- Suggestions for improvement.

NPM quality surveys

An NPM field trial review, an NPM main study review, and a data submission questionnaire enabled NPMs to self-report systematically on all key aspects of field operations and data submission. The NPM main study review made provision for NPMs to self-report on their:

- Use of *KeyQuest* for sampling and data entry;
- Translation, adaptation and verification procedures;
- Preparation of instruments;
- Implementation of exclusions standards; and
- Implementation of coding procedures.

The data submission questionnaire focused on matters specifically relating to the data, including the implementation of national and international options.

PISA test administrator test session report

A test session report for the recording of key test session information enabled the systematic monitoring of test administration. The test session report recorded data on:

- The session date and timing;
- The position of the test administrator;
- The conduct of the students; and
- The testing environment.

IMPLEMENTATION OF QUALITY MONITORING PROCEDURES

Milestone database

The consortium used project milestones negotiated individually with each NPM to monitor the progress of each national centre. Main study testing dates, national centre requirements, and consortium reporting imperatives provided the basis for timeline negotiation. Consortium experts used the milestone database to monitor the progress of national centres through key parts of the project and, when problems were identified, to advise on rectifying actions in order to forestall further operational problems.

National centre quality monitors – Site visits

A consortium representative visited most national centres in the two weeks prior to their main study. For some national centres it was not possible to visit before commencement of the main study due to international health and security alerts. In these cases the consortium representative visited at a time when the alerts were lifted. Consortium representatives visited all national centres. The NCQM used the visit to conduct a half-day PQM training session and a face-to-face interview with the NPM or their representative. Potential problems identified by the NCQM at a national centre were forwarded to the relevant consortium expert for appropriate action.

Video-conferencing facilities enabled the training of PQMs prior to the main study, where the site visit occurred after the main study had commenced.

A comprehensive knowledge of PISA operations and an extensive experience in PISA operations were the criteria for NCQM selection. The NCQMs were trained in conducting site visits to ensure their independence. Nationals with a formal association with the consortium did not visit their own national centre.

PISA quality monitors

NPMs nominated PQMs to the consortium. The candidate's formal independence from the national centre, their experience in, or familiarity with, school operations, their experience or familiarity with educational research, and an ability to speak English or French provided the basis for nomination and selection. Candidates nominated for PQM submitted a resume to the consortium. Where the resume did not match the selection criteria, further information or an alternate nomination was sought. In some countries where the PISA national centre was part of the ministry of education, and where there was a legislative requirement that all staff entering school be ministry employees, it was not possible to fulfil the criteria of PQM independence from the national centre. One national centre was not able to nominate candidates with the required criteria and in this case the consortium organised suitably qualified PQMs.

Typically, two PQMs were engaged for each country, with each PQM visiting seven or eight schools. An NCQM trained all PQMs. The NCQM and PQMs collaborated to develop a schedule of school visits, to ensure that a range of schools was covered and to ensure that the schedule of visits was both economically and practically feasible. The consortium paid the PQM expenses and fees.

The majority of school visits were unannounced. However, the need to organise transport and accommodation made it impractical to keep all PQM visits unannounced.

QUALITY MONITORING DATA

The quality-monitoring data collected from the quality-monitoring instrument was centralised in a single database. Data from the NCQMs, the PQMs, and the NPM quality surveys were data entered by the consortium. Consortium experts used consolidated quality-monitoring reports from the resulting central database to make country-by-country judgements on the quality of field operations, translation, school and student sampling, and coding. The consortium experts used the collected quality-monitoring information to cross check against their own records. The final reports by consortium experts were then used for the purpose of data adjudication (see Chapter 15).

An aggregated report on quality monitoring is also included as Appendix 9.

8

Survey Weighting and the Calculation of Sampling Variance

Survey weights were required to analyse PISA 2003 data, to calculate appropriate estimates of sampling error, and to make valid estimates and inferences. The consortium calculated survey weights for all assessed, ineligible and excluded students, and provided variables in the data that permit users to make approximately unbiased estimates of standard errors, to conduct significance tests and to create confidence intervals appropriately, given the sample design for PISA in each individual country.

SURVEY WEIGHTING

Students included in the final PISA sample for a given country are not all equally representative of the entire student population, despite random sampling of schools and students for selecting the sample. Survey weights must therefore be incorporated into the analysis.

There are several reasons why the survey weights are not the same for all students in a given country:

- A school sample design may intentionally over- or under-sample certain sectors of the school population: in the former case, so that they could be effectively analysed separately for national purposes, such as a relatively small but politically important province or region, or a sub-population using a particular language of instruction; and in the latter case, for reasons of cost, or other practical considerations,[1] such as very small or geographically remote schools.
- Information about school size available at the time of sampling may not have been completely accurate. If a school was expected to be very large, the selection probability was based on the assumption that only a sample of its students would be selected for PISA. But if the school turned out to be quite small, all students would have to be included and would have, overall, a higher probability of selection in the sample than planned, making these inclusion probabilities higher than those of most other students in the sample. Conversely, if a school thought to be small turned out to be large, the students included in the sample would have had smaller selection probabilities than others.
- School non-response, where no replacement school participated, may have occurred, leading to the under-representation of students from that kind of school, unless weighting adjustments were made. It is also possible that only part of the eligible population in a school (such as those 15-year-olds in a single grade) were represented by its student sample, which also requires weighting to compensate for the missing data from the omitted grades.
- Student non-response, within participating schools, occurred to varying extents. Students of the kind that could not be given achievement test scores (but were not excluded for linguistic or disability reasons) will be under-represented in the data unless weighting adjustments are made.
- Trimming weights to prevent undue influence of a relatively small subset of the school or student sample might have been necessary if a small group of students would otherwise have much larger weights than the remaining students in the country. This can lead to unstable estimates – large sampling errors – but cannot be estimated well. Trimming weights introduces a small bias into estimates, but greatly reduces standard errors.

The procedures used to derive the survey weights for PISA reflect the standards of best practice for analysing complex survey data, and the procedures used by the world's major statistical agencies. The same procedures were used in other international studies of educational achievement: the Third International Mathematics and Science Study (TIMSS), the Third International Mathematics and Science Study–Repeat (TIMSS-R), the Civic Education Study (CIVED), and the Progress in International Reading Literacy Study 2001 (PIRLS), which were all implemented by the International Association for the Evaluation of

Educational Achievement (IEA), and also in the International Assessment of Educational Progress (IAEP, 1991). (See Cochran, 1977 and Särndal *et al.,* 1992, for the underlying statistical theory on survey sampling texts.)

The weight, W_{ij}, for student j in school i consists of two base weights – the school and the within-school – and five adjustment factors, and can be expressed as:

$$W_{ij} = t_{2ij} f_{1i} f_{1ij}^{A} t_{1i} w_{2ij} w_{1i} \quad (8.1)$$

where:

- w_{1i}, the school base weight, is given as the reciprocal of the probability of inclusion of school i into the sample;
- w_{2ij}, the within-school base weight, is given as the reciprocal of the probability of selection of student j from within the selected school i;
- f_{1i} is an adjustment factor to compensate for non-participation by other schools that are somewhat similar in nature to school i (not already compensated for by the participation of replacement schools);
- f_{1ij}^{A} is an adjustment factor to compensate for the fact that, in some countries, in some schools only 15-year-old students who were enrolled in the modal grade for 15-year-olds were included in the assessment;
- t_{1i} is a school trimming factor, used to reduce unexpectedly large values of w_{1i}; and
- t_{2ij}, is a student trimming factor, used to reduce the weights of students with exceptionally large values for the product of all the preceding weight components.

The school base weight

The term w_{1i} is referred to as the school base weight. For the systematic probability proportional-to-size school sampling method used in PISA, this is given as:

$$w_{1i} = \begin{cases} int(g/i) \Big/ mos(i) & \text{if } mos(i) < int(g/i) \\ 1 & \text{otherwise} \end{cases} \quad (8.2)$$

The term $mos(i)$ denotes the measure of size given to each school on the sampling frame.

Despite country variations, $mos(i)$ was usually equal to the estimated number of 15-year-olds in the school, if it was greater than the predetermined target cluster size (35 in most countries).

If the enrolment of 15-year-olds was less than the Target Cluster Size (TCS), then $mos(i) = TCS$.

The term $int(g/i)$ denotes the sampling interval used within the explicit sampling stratum g that contains school i and is calculated as the total of $mos(i)$ values for all schools in stratum g, divided by the school sample size for that stratum.

Thus, if school i was estimated to have 100 15-year-olds at the time of sample selection, $mos(i) = 100$. If the country had a single explicit stratum (g=1) and the total of the values over all schools was 150 000, with a school sample size of 150, then $int(1/i) = 150000/150 = 1000$, for school i (and others in the sample), giving $w_{1i} = 1000/100 = 10.0$. Roughly speaking, the school can be thought of as representing about

10 schools from the population. In this example, any school with 1 000 or more 15-year-old students would be included in the sample with certainty, with a base weight of $w_{1i} = 1$.

The school weight trimming factor

Once school base weights were established for each sampled school in the country, verifications were made separately within each explicit sampling stratum to see if the school weights required trimming. The school trimming factor t_{1i}, is the ratio of the trimmed to the untrimmed school base weight, and is equal to 1.0000 for most schools and therefore most students, and never exceeds this value. (See Table 8.1 for the number of school records in each country that received some kind of base weight trimming.)

The school-level trimming adjustment was applied to schools that turned out to be much larger than was believed at the time of sampling – where 15-year-old enrolment exceeded $3 \times \max\left(TCS,\ mos\,(i)\right)$. For example, if TCS = 35, then a school flagged for trimming had more than 105 PISA-eligible students, and more than three times as many students as was indicated on the school sampling frame. Because the student sample size was set at TCS regardless of the actual enrolment, the student sampling rate was much lower than anticipated during the school sampling. This meant that the weights for the sampled students in these schools would have been more than three times greater than anticipated when the school sample was selected. These schools had their school base weights trimmed by having $mos\,(i)$ replaced by $3 \times \max\left(TCS,\ mos\,(i)\right)$ in the school base weight formula.

The student base weight

The term w_{2ij} is referred to as the student base weight, which with the PISA procedure for sampling students, did not vary across students (j) within a particular school i. This is given as:

$$w_{2ij} = {enr\,(i)} \Big/ {sam\,(i)} \qquad (8.3)$$

where $enr\,(i)$ is the actual enrolment of 15-year-olds in the school (and so, in general, is somewhat different from the estimated $mos\,(i)$), and $sam\,(i)$ is the sample size within school i. It follows that if all students from the school were selected, then w_{2ij} = 1 for all eligible students in the school. For all other cases $w_{2ij} > 1$.

School non-response adjustment

In order to adjust for the fact that those schools that declined to participate, and were not replaced by a replacement school, were not in general typical of the schools in the sample as a whole, school-level non-response adjustments were made. Several groups of somewhat similar schools were formed within a country, and within each group the weights of the responding schools were adjusted to compensate for the missing schools and their students. The compositions of the non-response groups varied from country to country, but were based on cross-classifying the explicit and implicit stratification variables used at the time of school sample selection. Usually, about 10 to 15 such groups were formed within a given country, depending upon school distribution with respect to stratification variables. If a country provided no implicit stratification variables, schools were divided into three roughly equal groups, within each stratum, based on their size (small, medium or large). It was desirable to ensure that each group had at least six participating schools, as small groups can lead to unstable weight adjustments, which in turn would inflate the sampling variances. However, it was not necessary to collapse cells where all schools participated, as the school non-response adjustment factor was 1.0 regardless of whether cells were collapsed or not. Adjustments

greater than 2.0 were flagged for review, as they can cause increased variability in the weights, and lead to an increase in sampling variances. In either of these situations, cells were generally collapsed over the last implicit stratification variable(s) until the violations no longer existed. In countries with very high overall levels of school non-response after school replacement, the requirement for school non-response adjustment factors all to be below 2.0 was waived.

Within the school non-response adjustment group containing school i, the non-response adjustment factor was calculated as:

$$f_{1i} = \frac{\sum_{k \in \Omega(i)} w_{1k}\, enr(k)}{\sum_{k \in \Gamma(i)} w_{1k}\, enr(k)} \qquad (8.4)$$

where the sum in the denominator is over $\Gamma(i)$, the schools within the group (originals and replacements) that participated, while the sum in the numerator is over $\Omega(i)$, those same schools, plus the original sample schools that refused and were not replaced. The numerator estimates the population of 15-year-olds in the group, while the denominator gives the size of the population of 15-year-olds directly represented by participating schools. The school non-response adjustment factor ensures that participating schools are weighted to represent all students in the group. If a school did not participate because it had no eligible students enrolled, no adjustment was necessary since this was neither non-response nor under-coverage.

Table 8.1 shows the number of school non-response classes that were formed for each country, and the variables that were used to create the cells.

Grade non-response adjustment

In two countries (Denmark and the United States), several schools agreed to participate in PISA, but required that participation be restricted to 15-year-olds in the modal grade for 15-year-olds, rather than all 15-year-olds, because of perceived administrative inconvenience. Since the modal grade generally included the majority of the population to be covered, some of these schools were accepted as participants. For the part of the 15-year-old population in the modal grade, these schools were respondents, while for the rest of the grades in the school with 15-year-olds, this school was a refusal. This situation occasionally arose for a grade other than the modal grade because of other reasons, such as other testing being carried out for certain grades at the same time as the PISA assessment. To account for this, a special non-response adjustment was calculated at the school level for students not in the modal grade (and was automatically 1.0 for all students in the modal grade).

Within the same non-response adjustment groups used for creating school non-response adjustment factors, the grade non-response adjustment factor for all students in school i, f_{1i}^{A}, is given as:

$$f_{1i}^{A} = \begin{cases} \sum_{k \in C(i)} w_{1k}\, enra(k) \Big/ \sum_{k \in B(i)} w_{1k}\, enra(k) & \text{for students not in the modal grade} \\ 1 & \text{otherwise} \end{cases} \qquad (8.5)$$

The variable *enra(k)* is the approximate number of 15-year-old students in school k but not in the modal grade. The set B(i) is all schools that participated for all eligible grades (from within the non-response adjustment group with school (i)), while the set C(i) includes these schools and those that only participated for the modal responding grade.

Table 8.1. ■ **Non-response classes**

	Implicit stratification variables used to create school non-response cells (within explicit stratum), and number of original and final cells	Number of original cells	Number of final cells
Australia	Urban/rural (2)	46	30
Austria	Size (large/small)	30	28
Belgium	Flanders – school proportion of overage students (continuous); French Community – school size (3), school proportion of overage students (continuous); German Community – school type (3), school size (4)	222	46
Brazil	School type (3), urban/rural (2), index of school infrastructure (4)	51	38
Canada	Public/private (2), urban/rural (2)	165	71
Czech Republic	Regions (14) for four school types	140	135
Denmark	School type (4), county (15)	44	18
Finland	Size (3)	35	35
France	Size (3)	18	10
Germany	School type (5) for normal school, for state (16), for vocational schools	67	37
Greece	School type (4), public/private	30	13
Hong Kong-China	For strata 1 and 2, academic intake (3), for independent schools (stratum 3) local or international funding (2)	8	7
Hungary	geographic region (7+1 for missing) for strata 1-4, for stratum 5, TIMSS explicit (TIMSS population variable with two levels) and implicit (20 regions and three levels of urbanization) stratifiers	87	43
Iceland	Urban/rural, school size (4)	33	30
Indonesia	School type (5), public/private (2), national achievement score categories (3)	202	190
Ireland	School type (3), school gender composition categories (5)	24	13
Italy	Public/private (2)	74	30
Japan	Levels (4) of proportions of students taking university or college entrance exams	15	13
Korea	School level (2)	11	10
Latvia	Urbanicity (3), school type (3)	20	8
Liechtenstein	None, three cells formed based on sizes	3	3
Luxembourg	Size (3)	10	4
Macao-China	Size classes (3) for strata 2 and 3	7	7
Mexico	School type (6), urban/rural (2), school level (3), program (3 or 4 depending on school level)	299	259
Netherlands	School type (6)	10	6
New Zealand	Public/private (2), socio-economic status category (3), urban/rural (2)	11	9
Norway	Size (3)	12	7
Poland	Urbanicity (4)	7	5
Portugal	Public/private (2), socio-economic status category (4)	28	20
Russian Federation	School type (3), urbanicity (5) [no school non-response adjustments]	169	157
Serbia	Urban/rural, school type (7), Hungarian students or not	68	64
Slovak Republic	School type (9), language (2), authority (9)	89	53
Spain	For Catalonia: size of town (3), province (numerous); for other regions: province (numerous)	107	107
Sweden	School level (2), income quartile (4), responsible authority(2), urbanicity (5), geographic area (many) –various combinations of these depending on explicit stratum	45	20
Switzerland	School type (many levels), canton (many levels)	171	84
Thailand	Region (13)	58	39
Tunisia	Levels of grade repetition for three school levels (numerous)	41	14
Turkey	School type (18)	123	112
United Kingdom	England – school type (3), exam grade (7), gender (3), region (4, derived from 150 levels of LEA); Wales – secondary/independent, exam grade (4) for secondary schools; Northern Ireland – school type (3), exam grade bands (7), region (5); Scotland – school size (3).	116	47
United States	Gradprop (5), public/private (2), region (4), urbanicity (8), minstat (2)	172	39
Uruguay	Program type (3-7 levels depending on explicit stratum), shift (4 or 5 depending on program, for several strata and are for another stratum), area (3) for one stratum	63	45

© OECD 2005

This procedure gave, for each school, a single grade non-response adjustment factor that depended upon its non-response adjustment class. Each individual student received this factor value if they did not belong to the modal grade, and 1.0000 if they belonged to the modal grade. In general, this factor is not the same for all students within the same school.

Student non-response adjustment

Within each participating school and high/low grade combination, the student non-response adjustment f_{2i} was calculated as:

$$f_{2i} = \frac{\sum_{k \in X(i)} f_{1i} w_{1i} w_{2ik}}{\sum_{k \in \Delta(i)} f_{1i} w_{1i} w_{2ik}} \qquad (8.6)$$

where the set $\Delta(i)$ is all assessed students in the school / grade combination and the set $X(i)$ is all assessed students in the school / grade combination plus all others who should have been assessed (*i.e.* who were absent, but not excluded or ineligible). The high and low grade categories in each country were defined so as to each contain a substantial proportion of the PISA population.

In most cases, this student non-response factor reduces the ratio of the number of students who should have been assessed to the number who were assessed. In some cases of small cells (*i.e.* school/grade category combinations) sizes (fewer than ten respondents), it was necessary to collapse cells together, and then the more complex formula above applied. Additionally, an adjustment factor greater than 2.0 was not allowed for the same reasons noted under school non-response adjustments. If this occurred, the cell with the large adjustment was collapsed with the closest cell in the same school non-response cell.

Some schools in some countries had very low student response levels. In these cases it was determined that the small sample of assessed students was potentially too biased as a representation of the school to be included in the PISA data. For any school where the student response rate was below 25 per cent, the school was therefore treated as a non-respondent, and its student data were removed. In schools with between 25 and 50 per cent student response, the student non-response adjustment described above would have resulted in an adjustment factor of between 2.0000 and 4.0000, and so these schools were collapsed with others to create student non-response adjustments.[2]

Trimming student weights

This final trimming check was used to detect student records that were unusually large compared to those of other students within the same explicit stratum. The sample design was intended to give all students from within the same explicit stratum an equal probability of selection and therefore equal weight, in the absence of school and student non-response. As already noted, poor prior information about the number of eligible students in each school could lead to substantial violations of this principle. Moreover, school, grade and student non-response adjustments, as well as, occasionally, inappropriate student sampling could in a few cases accumulate to give a few students in the data relatively large weights, which adds considerably to sampling variance. The weights of individual students were therefore reviewed, and where the weight was more than four times the median weight of students from the same explicit sampling stratum, it was trimmed to be equal to four times the median weight for that explicit stratum.

The student trimming factor, t_{2ij}, is equal to the ratio of the final student weight to the student weight adjusted for student non-response, and therefore equal to 1.0000 for the great majority of students. The final weight variable on the data file was called *w_fstuwt*, which is the final student weight that incorporates any student-level trimming. Table 8.2 shows the number of students with weights trimmed at this point in the process (*i.e.* $t_{2ij} < 1.0000$) for each country and the number of schools for which the school base weight was trimmed (*i.e.* $t_{1i} < 1.0000$).

CALCULATING SAMPLING VARIANCE

To estimate the sampling variances of PISA estimates, a replication methodology was employed. This reflected the variance in estimates due to the sampling of schools and students. Additional variance due to the use of plausible values from the posterior distributions of scaled scores was captured separately, although computationally the two components can be carried out in a single program, such as WesVar 4 (Westat, 2000).

The balanced repeated replication variance estimator

The approach used for calculating sampling variances for PISA is known as Balanced Repeated Replication (BRR), or Balanced Half-Samples; the particular variant known as Fay's method was used. This method is very similar in nature to the Jackknife method used in previous international studies of educational achievement, such as TIMSS, and it is well documented in the survey sampling literature (Rust, 1985; Rust and Rao, 1996; Shao, 1996; Wolter, 1985). The major advantage of BRR over the Jackknife is that the Jackknife method is not fully appropriate for use with non-differentiable functions of the survey data, most noticeably quantiles. It provides unbiased estimates, but not consistent ones. This means that, depending upon the sample design, the variance estimator can be very unstable, and despite empirical evidence that it can behave

Table 8.2 ■ **School and student trimming**

Country	Number of schools trimmed	Number of students trimmed
Australia	1	0
Austria	0	0
Belgium	0	0
Belgium-Flanders	0	0
Belgium-French	0	0
Belgium-German	0	0
Brazil	0	0
Canada	0	0
Czech Republic	0	0
Denmark	0	0
Finland	0	0
France	0	0
Germany	0	0
Greece	0	0
Hong Kong-China	1	0
Hungary	0	6
Iceland	0	0
Indonesia	5	0
Ireland	0	0
Italy	0	0
Japan	0	0
Korea	0	0
Latvia	0	0
Liechtenstein	0	0
Luxembourg	0	0
Macao-China	0	35
Mexico	0	107
Netherlands	5	0
New Zealand	0	0
Norway	0	0
Poland	0	0
Portugal	1	0
Russian Federation	11	0
Serbia	0	0
Slovak Republic	0	0
Spain	0	0
Sweden	1	0
Switzerland	0	91
Thailand	0	0
Tunisia	0	0
Turkey	1	0
United Kingdom	2	0
England	1	0
Northern Ireland	1	0
Wales	0	0
Scotland	0	0
United States	2	0
Uruguay	0	0

well in a PISA-like design, theory is lacking. In contrast, BRR does not have this theoretical flaw. The standard BRR procedure can become unstable when used to analyse sparse population subgroups, but Fay's modification overcomes this difficulty, and is well justified in the literature (Judkins, 1990).

The BRR approach was implemented as follows, for a country where the student sample was selected from a sample of, rather than all, schools:

- Schools were paired on the basis of the explicit and implicit stratification and frame ordering used in sampling. The pairs were originally sampled schools, or pairs that included a participating replacement if an original refused. For an odd number of schools within a stratum, a triple was formed consisting of the last school and the pair preceding it.
- Pairs were numbered sequentially, 1 to H, with pair number denoted by the subscript h. Other studies and the literature refer to such pairs as variance strata or zones, or pseudo-strata.
- Within each variance stratum, one school (the primary sampling unit, PSU) was randomly numbered as 1, the other as 2 (and the third as 3, in a triple), which defined the variance unit of the school. Subscript j refers to this numbering.
- These variance strata and variance units (1, 2, 3) assigned at school level are attached to the data for the sampled students within the corresponding school.
- Let the estimate of a given statistic from the full student sample be denoted as X^*. This is calculated using the full sample weights.
- A set of 80 replicate estimates, X_t^* (where t runs from 1 to 80), was created. Each of these replicate estimates was formed by multiplying the sampling weights from one of the two PSUs in each stratum by 1.5, and the weights from the remaining PSUs by 0.5. The determination as to which PSUs received inflated weights, and which received deflated weights, was carried out in a systematic fashion, based on the entries in a Hadamard matrix of order 80. A Hadamard matrix contains entries that are +1 and –1 in value, and has the property that the matrix, multiplied by its transpose, gives the identity matrix of order 80, multiplied by a factor of 80. (Examples of Hadamard matrices are given in Wolter, 1985.)
- In cases where there were three units in a triple, either one of the schools (designated at random) received a factor of 1.7071 for a given replicate, with the other two schools receiving factors of 0.6464, or else the one school received a factor of 0.2929 and the other two schools received factors of 1.3536. The explanation of how these particular factors came to be used is explained in Appendix 12 of the *PISA 2000 Technical Report* (OECD, 2002).
- To use a Hadamard matrix of order 80 requires that there be no more than 80 variance strata within a country, or else that some combining of variance strata be carried out prior to assigning the replication factors via the Hadamard matrix. The combining of variance strata does not cause any bias in variance estimation, provided that it is carried out in such a way that the assignment of variance units is independent from one stratum to another within strata that are combined. That is, the assignment of variance units must be completed before the combining of variance strata takes place. This approach was used for PISA.
- The reliability of variance estimates for important population subgroups is enhanced if any combining of variance strata that is required is conducted by combining variance strata from different subgroups. Thus in PISA, variance strata that were combined were selected from different explicit sampling strata and, to the extent possible, from different implicit sampling strata also.

- In some countries, it was not the case that the entire sample was a two-stage design, of first sampling schools and then sampling students. In some countries for part of the sample (and for the entire samples for Iceland, Macao-China, Liechtenstein and Luxembourg), schools were included with certainty into the sampling, so that only a single stage of student sampling was carried out for this part of the sample. In these cases instead of pairing schools, pairs of individual students were formed from within the same school (and if the school had an odd number of sampled students, a triple of students was formed also). The procedure of assigning variance units and replicate weight factors was then conducted at the student level, rather than at the school level.
- In contrast, in a few countries there was a stage of sampling that preceded the selection of schools, for at least part of the sample. This was done in a major way in the Russian Federation and Turkey. In these cases there was a stage of sampling that took place before the schools were selected. Then the procedure for assigning variance strata, variance units and replicate factors was applied at this higher level of sampling. The schools and students then inherited the assignment from the higher-level unit in which they were located.
- The variance estimator is then:

$$V_{BRR}\left(X^*\right)=0.05\sum_{t=1}^{80}\left\{\left(X_t^*-X^*\right)^2\right\} \qquad (8.7)$$

The properties of BRR have been established by demonstrating that it is unbiased and consistent for simple linear estimators (*i.e.* means from straightforward sample designs), and that it has desirable asymptotic consistency for a wide variety of estimators under complex designs, and through empirical simulation studies.

Reflecting weighting adjustments

This description glosses over one aspect of the implementation of the BRR method. Weights for a given replicate are obtained by applying the adjustment to the weight components that reflect selection probabilities (the school base weight in most cases), and then re-computing the non-response adjustment replicate by replicate.

Implementing this approach required that the consortium produce a set of replicate weights in addition to the full sample weight. Eighty such replicate weights were needed for each student in the data file. The school and student non-response adjustments had to be repeated for each set of replicate weights.

To estimate sampling errors correctly, the analyst must use the variance estimation formula above, by deriving estimates using the t-th set of replicate weights instead of the full sample weight. Because of the weight adjustments (and the presence of occasional triples), this does not mean merely increasing the final full sample weights for half the schools by a factor of 1.5 and decreasing the weights from the remaining schools by a factor of 0.5. Many replicate weights will also be slightly disturbed, beyond these adjustments, as a result of repeating the non-response adjustments separately by replicate.

Formation of variance strata

With the approach described above, all original sampled schools were sorted in stratum order (including refusals, excluded and ineligible schools) and paired, by contrast to other international education assessments such TIMSS and TIMSS-R that have paired participating schools only. However, these studies did not use an approach reflecting the impact of non-response adjustments on sampling variance. This is unlikely to be a big component of variance in any PISA country, but the procedure gives a more accurate estimate of sampling variance.

Countries where all students were selected for PISA

In Iceland, Liechtenstein and Luxembourg, all eligible students were selected for PISA. It might be considered surprising that the PISA data should reflect any sampling variance in these countries, but students have been assigned to variance strata and variance units, and the BRR formula does give a positive estimate of sampling variance for three reasons. First, in each country there was some student non-response, and, in the case of Iceland and Luxembourg, some school non-response. Not all eligible students were assessed, giving sampling variance. Second, only 55 per cent of the students were assessed in reading and science. Third, the issue is to make inference about educational systems and not particular groups of individual students, so it is appropriate that a part of the sampling variance reflect random variation between student populations, even if they were to be subjected to identical educational experiences. This is consistent with the approach that is generally used whenever survey data are used to try to make direct or indirect inference about some underlying system.

Notes

1 Note that this is not the same as excluding certain portions of the school population. This also happened in some cases, but cannot be addressed adequately through the use of survey weights.

2 Chapter 12 describes these schools as being treated as non-respondents for the purpose of response rate calculation, even though their student data were used in the analyses.

9

Scaling PISA Cognitive Data

The mixed co-efficients multinomial logit model as described by Adams *et al.* (1997) was used to scale the PISA data, and implemented by *ConQuest* software (Wu *et al.,* 1997).

THE MIXED CO-EFFICIENTS MULTINOMIAL LOGIT MODEL

The model applied to PISA is a generalised form of the Rasch model. The model is a mixed co-efficients model where items are described by a fixed set of unknown parameters, ξ, while the student outcome levels (the latent variable), θ, is a random effect.

Assume that I items are indexed $i = 1,\ldots,I$ with each item admitting $K_i + 1$ response categories indexed $k = 0,1,\ldots,K_i$. Use the vector valued random variable $\mathrm{X}_i = \left(X_{i1}, X_{i2},\ldots, X_{iK_i}\right)^T$, where

$$X_{ij} = \begin{cases} 1 & \text{if response to item } i \text{ is in category } j \\ 0 & \text{otherwise} \end{cases}, \qquad (9.1)$$

to indicate the $K_i + 1$ possible responses to item i.

A vector of zeroes denotes a response in category zero, making the zero category a reference category, which is necessary for model identification. Using this as the reference category is arbitrary, and does not affect the generality of the model. The X_i can also be collected together into the single vector $\mathrm{X}^T = \left(\mathrm{X}_1^T, \mathrm{X}_2^T,\ldots,\mathrm{X}_I^T\right)$, called the response vector (or pattern). Particular instances of each of these random variables are indicated by their lower case equivalents; x, $\mathrm{x_i}$ and x_{ik}.

Items are described through a vector $\xi^T = \left(\xi_1,\xi_2,\ldots,\xi_p\right)$, of p parameters. Linear combinations of these are used in the response probability model to describe the empirical characteristics of the response categories of each item. D Design vectors $\mathrm{a}_{ij}, \left(i = 1,\ldots,I;\ j = 1,\ldots K_i\right)$, each of length p, which can be collected to form a design matrix $\mathrm{A}^T = \left(\mathrm{a}_{11},\mathrm{a}_{12},\ldots,\mathrm{a}_{1K_1},\mathrm{a}_{21},\ldots,\mathrm{a}_{2K_2},\ldots,\mathrm{a}_{IK_I}\right)$ define these linear combinations.

The multi-dimensional form of the model assumes that a set of D traits underlies the individuals' responses. The D latent traits define a D-dimensional latent space. The vector $\theta = \left(\theta_1,\theta_2,\ldots,\theta_D\right)$, represents an individual's position in the D-dimensional latent space.

The model also introduces a scoring function that allows specifying the score or performance level assigned to each possible response category to each item. To do so, the notion of a response score b_{ijd} is introduced, which gives the performance level of an observed response in category j, item i, dimension d. The scores across D dimensions can be collected into a column vector $\mathrm{b}_{ik} = \left(b_{ik1}, b_{ik2},\ldots,b_{ikD}\right)^T$ and again collected into the scoring sub-matrix for item i, $\mathrm{B}_i = \left(\mathrm{b}_{i1},\mathrm{b}_{i2},\ldots,\mathrm{b}_{iD}\right)^T$ and then into a scoring matrix $\mathrm{B} = \left(\mathrm{B}_1^T,\mathrm{B}_2^T,\ldots,\mathrm{B}_I^T\right)^T$ for the entire test. (The score for a response in the zero category is zero, but other responses may also be scored zero).

The probability of a response in category j of item i is modelled as

$$\Pr\left(\mathrm{X}_{ij} = 1; \mathrm{A}, \mathrm{B}, \xi \mid \theta\right) = \frac{\exp\left(\mathrm{b}_{ij}\theta + \mathrm{a}'_{ij}\xi\right)}{\sum_{k=1}^{K_i}\exp\left(\mathrm{b}_{ik}\theta + \mathrm{a}'_{ik}\xi\right)}. \qquad (9.2)$$

For a response vector we have

$$f(x;\xi|\theta) = \Psi(\theta,\xi)\exp\left[x'(B\theta + A\xi)\right], \qquad (9.3)$$

with

$$\Psi(\theta,\xi) = \left\{\sum_{z\in\Omega}\exp\left[z^T(B\theta + A\xi)\right]\right\}^{-1} \qquad (9.4)$$

where Ω is the set of all possible response vectors.

The population model

The item response model is a conditional model, in the sense that it describes the process of generating item responses conditional on the latent variable, θ. The complete definition of the model, therefore, requires the specification of a density, $f_\theta(\theta;\alpha)$ for the latent variable, θ. Let α symbolise a set of parameters that characterise the distribution of θ. The most common practice, when specifying uni-dimensional marginal item response models, is to assume that students have been sampled from a normal population with mean μ and variance σ^2. That is:

$$f_\theta(\theta;\alpha) \equiv f_\theta(\theta;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \qquad (9.5)$$

or equivalently

$$\theta = \mu + E \qquad (9.6)$$

where $E \sim N(0,\sigma^2)$.

Adams *et al.* (1997) discuss how a natural extension of (9.6) is to replace the mean, μ with the regression model, $Y_n^T\rceil$ where Y_n is a vector of *u,* fixed and known values for student *n,* and β is the corresponding vector of regression co-efficients. For example, Y_n could be constituted of student variables such as gender or socio-economic status. Then the population model for student *n*, becomes,

$$\theta_n = Y_n^T\rceil + E_n \qquad (9.7)$$

where it is assumed that the E_n are independently and identically normally distributed with mean zero and variance σ^2 so that (9.7) is equivalent to:

$$f_\theta(\theta_n;Y_n,b,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{1}{2\sigma^2}(\theta_n - Y_n^T\rceil)^T(\theta_n - Y_n^T\rceil)\right], \qquad (9.8)$$

a normal distribution with mean $Y_n^T\rceil$ and variance σ^2. If (9.8) is used as the population model then the parameters to be estimated are β, σ^2 and ξ.

The generalisation needs to be taken one step further to apply it to the vector valued θ rather than the scalar valued θ. The extension results in the multivariate population model:

$$f_\theta(\theta_n;W_n,\gamma,\Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left[-\frac{1}{2}(\theta_n - \gamma W_n)^T\Sigma^{-1}(\theta_n - \gamma W_n)\right], \qquad (9.9)$$

where γ is a $u\times d$ matrix of regression co-efficients, Σ is a $d\times d$ variance-covariance matrix and W_n is a $u\times 1$ vector of fixed variables.

In PISA, the W_n variables are referred to as conditioning variables.

Combined model

In (9.10), the conditional item response model (9.4) and the population model (9.9) are combined to obtain the unconditional, or marginal, item response model:

$$f_x(x;\xi,\gamma,\Sigma)=\int_\theta f_x(x;\xi\,|\theta)\,f_\theta(\theta;\gamma,\Sigma)\,d\,\theta \qquad (9.10)$$

It is important to recognise that under this model the locations of individuals on the latent variables are not estimated. The parameters of the model are γ, Σ and ξ.

The procedures used to estimate model parameters are described in Adams *et al.* (1997a), Adams *et al.* (1997b), and Wu *et al.* (1997).

For each individual it is possible however to specify a posterior distribution for the latent variable, given by:

$$\begin{aligned} h_\theta(\theta_n;W_n,\xi,\gamma,\Sigma\,|x_n) &= \frac{f_x(x_n;\xi\,|\theta_n)\,f_\theta(\theta_n;W_n,\gamma,\Sigma)}{f_x(x_n;W_n,\xi,\gamma,\Sigma)} \\ &= \frac{f_x(x_n;\xi\,|\theta_n)\,f_\theta(\theta_n;W_n,\gamma,\Sigma)}{\int_{\theta_n} f_x(x_n;\xi\,|\theta_n)\,f_\theta(\theta_n;W_n,\gamma,\Sigma)} \end{aligned} \qquad (9.11)$$

APPLICATION TO PISA

In PISA, this model was used in three steps: national calibrations, international scaling and student score generation.

For both the national calibrations and the international scaling, the conditional item response model (9.3) is used in conjunction with the population model (9.9), but conditioning variables are not used. That is, it is assumed that students have been sampled from a multivariate normal distribution.

In PISA 2003 the main scaling model was seven-dimensional, made up of one reading, one science, one problem solving and four mathematics dimensions. The design matrix was chosen so that the partial credit model was used for items with multiple score categories and the simple logistic model was fit to the dichotomously scored items.

National calibrations

National calibrations were performed separately country-by-country using unweighted data. The results of these analyses, which were used to monitor the quality of the data and to make decisions regarding national item treatment, are given in Chapter 13.

The outcomes of the national calibrations were used to make a decision about how to treat each item in each country. This means that: an item may be deleted from PISA altogether if it has poor psychometric characteristics in more than ten countries (a "dodgy" item); it may be regarded as not-administered in particular countries if it has poor psychometric characteristics in those countries but functions well in the vast majority of others; or an item with sound characteristics in each country but which shows substantial item-by-country interactions may be regarded as a different item (for scaling purposes) in each country (or in some subset of countries) that is, the difficulty parameter will be free to vary across countries. Both

the second and third options have the same impact on comparisons between countries. That is, if an item is identified as behaving differently in different countries, choosing either the second or third option will have the same impact on inter-country comparisons. The choice between them could, however, influence within-country comparisons.

When reviewing the national calibrations, particular attention was paid to the fit of the items to the scaling model, item discrimination and item-by-country interactions.

Item response model fit (infit mean square)

For each item parameter, the *ConQuest* fit mean square statistic index (Wu *et al.,* 1997) was used to provide an indication of the compatibility of the model and the data. For each student, the model describes the probability of obtaining the different item scores. It is therefore possible to compare the model prediction and what has been observed for one item across students. Accumulating comparisons across cases gives us an item-fit statistic. As the fit statistics compare an observed value with a predicted value, the fit is an analysis of residuals. In the case of the item infit mean square, values near one are desirable. An infit mean square greater than one is often associated with a low discrimination index, and an infit mean square less than one is often associated with a high discrimination index.

Discrimination co-efficients

For each item, the correlation between the students' score and aggregate score on the set for the same domain and booklet as the item of interest was used as an index of discrimination. If p_{ij} $(= x_{ij} / m_i)$ is the proportion of score levels that student i achieved on item j, and $p_i = \sum_j P_{ij}$, (where the summation is of the items from the same booklet and domain as item j) is the sum of the proportions of the maximum score achieved by student i, then the discrimination is calculated as the product-moment correlation between p_{ij} and p_i for all students. For multiple-choice and short-answer items, this index will be the usual point-biserial index of discrimination.

The point-biserial index of discrimination for a particular category of an item is a comparison of the aggregate score between students selecting that category and all other students. If the category is the correct answer, the point-biserial index of discrimination should be higher than 0.25. Non-key categories should have a negative point-biserial index of discrimination. The point-biserial index of discrimination for a partial credit item should be ordered, *i.e.* categories scored 0 should be lower than the point-biserial correlation of categories scored 1, and so on.

Item-by-country interaction

The national scaling provides nationally specific item parameter estimates. The consistency of item parameter estimates across countries was of particular interest. If the test measured the same latent trait per domain in all countries, then items should have the same relative difficulty, or, more precisely, would fall within the interval defined by the standard error on the item parameter estimate.

National reports

After national scaling, five reports were returned to each participating country to assist in reviewing their data with the consortium:

- *Report 1* presented the results of a basic item analysis in tabular form. For each item, the number of students, the percentage of students, the point-biserial correlation, and student-centred Item Response Theory (IRT) ability average were provided for each valid category.
- *Report 2* provided, for each item and for each valid category, the point-biserial correlation and the student-centred IRT ability average in graphical form.
- *Report 3* provided a graphical comparison of the item infit mean square co-efficients and the item discrimination co-efficients computed at national and international levels.
- *Report 4* provided a graphical comparison of both the item difficulty parameter and the item thresholds, computed at national and international levels.
- *Report 5* listed the items that national project managers (NPMs) needed to check for mistranslation and/or misprinting, referred to as dodgy items.
- *Report 6* provides in a graphical form a comparison of the deviation of observed scores from expected scores for each item.

Report 1: Descriptive statistics on individual items in tabular form

A detailed item-by-item report was provided in tabular form showing the basic item analysis statistics at the national level (see Figure 9.1).

The table shows each possible response category for each item. The second column indicates the score assigned to the different categories. For each category, the number and percentage of students responding is shown, along with the point-biserial correlation and the associated *t* statistic. Note that for the item in the example the correct answer is '4', indicated by the '1' in the score column; thus the point-biserial for a response of '4' is the item's discrimination index, also shown along the top. The two last columns, *PV1Avg:1* and *PV1 SD:1*, show the average ability of students responding in each category and the standard deviation

Figure 9.1 ■ **Example of item statistics shown in Report 1**

```
Item 1
------
item:1 (M033Q01)
Cases for this item   1258   Discrimination  0.27
Item Threshold(s)    -2.06   Weighted MNSQ   1.11
--------------------------------------------------------------------------------
 Label    Score     Count   % of tot  Pt Bis     t  (p)   PV1Avg:1 PV1 SD:1
--------------------------------------------------------------------------------
   0                   0       0.00     NA       NA (.000)    NA       NA
   1       0.00        8       0.64   -0.07    -2.32(.021) -0.67     1.25
   2       0.00       76       6.04   -0.15    -5.46(.000) -0.62     1.13
   3       0.00       94       7.47   -0.18    -6.47(.000) -0.64     1.12
   4       1.00     1069      84.98    0.27     9.91(.000)  0.17     1.12
   5                   0       0.00     NA       NA (.000)    NA       NA
   6                   0       0.00     NA       NA (.000)    NA       NA
   7                   0       0.00     NA       NA (.000)    NA       NA
   8       0.00        4       0.32   -0.06    -2.31(.021) -1.04     1.42
   9       0.00        7       0.56   -0.05    -1.88(.060) -0.66     1.21
================================================================================
```

for it. The average ability is calculated by domain. In this example the average ability of those students who responded correctly (category 4) is 0.17, while the average ability of those students who responded incorrectly (categories 1, 2, 3) is around -0.6.

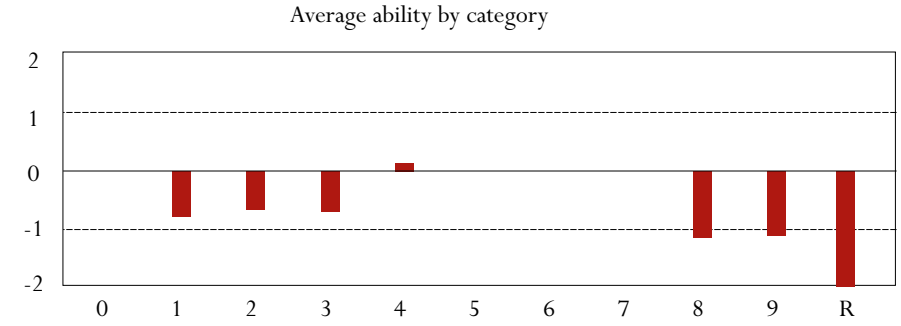
Report 2: Descriptive statistics on individual items in graphical form

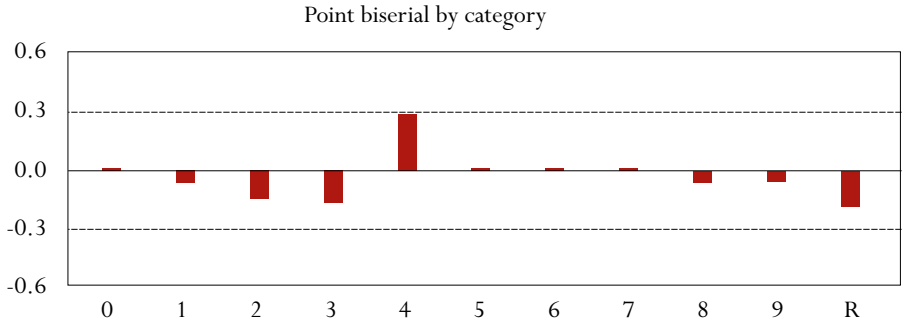
Report 2 (see Figure 9.2) graphs the ability average and the point-biserial correlation by category. Average ability by category is calculated by domain and centred for each item. This makes it easy to identify positive and negative ability categories, so that checks can be made to ensure that, for multiple-choice items, the key category has the highest average ability estimate, and for constructed-response items, the mean abilities are ordered consistently with the score levels. The displayed graphs also facilitate the process of identifying the following anomalies:

- A non-key category with a positive point-biserial or a point-biserial higher than the key category;
- A key category with a negative point-biserial; and
- For partial-credit items, average abilities (and point-biserials) not increasing with the score points.

Figure 9.2 ■ **Example of item statistics shown in Report 2**

Students	0	8	76	94	1069	0	0	0	4	7	18
%	0	1	6	7	84	0	0	0	0	1	1

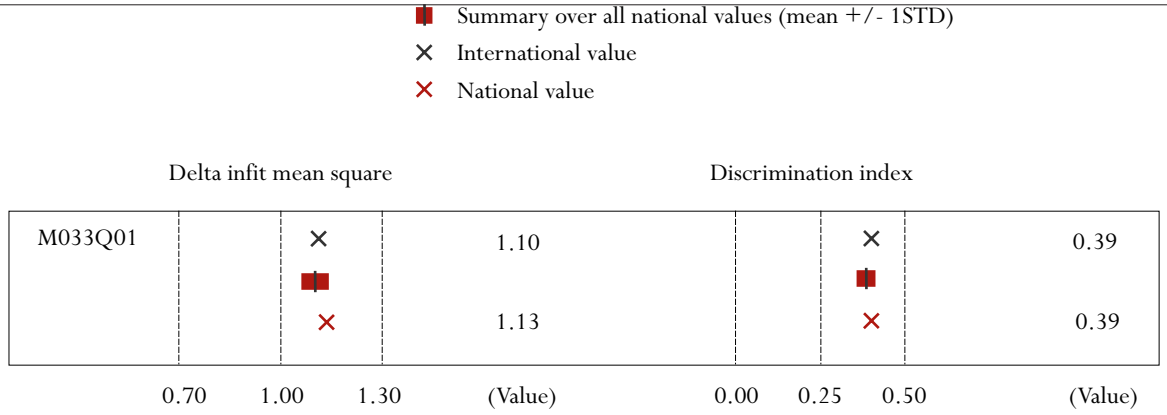




ID: M033Q01
Name: View room Q1

Discrimination: 0.27
Key: 4

Figure 9.3 ■ **Example of item statistics shown in Report 3**



Report 3: Comparison of national and international infit mean square and discrimination co-efficients

The national scaling provided the infit mean square, the point-biserial correlation, the item parameter estimate (or difficulty estimate) and the thresholds for each item in each country. Reports 3 and 4 (see Figures 9.3 and 9.4) compare the value computed for one country with those computed for all other countries and with the value computed at international level for each item.

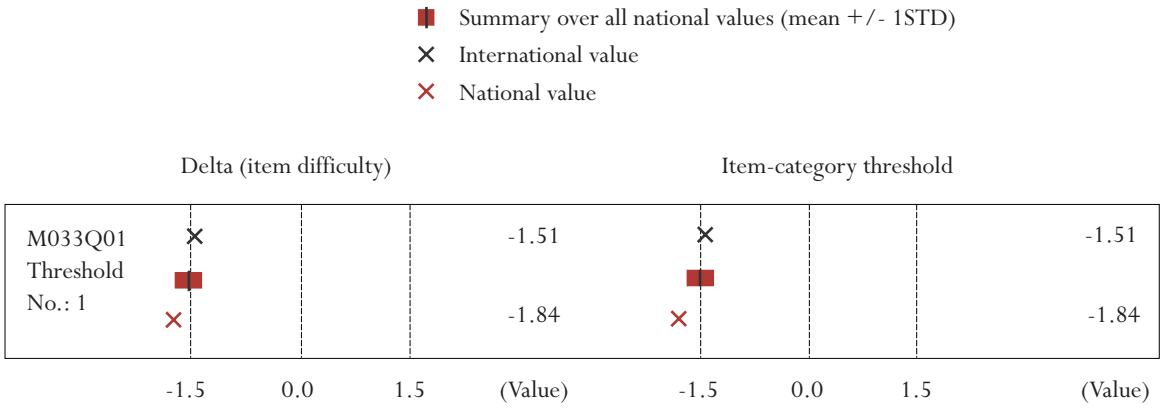
The black crosses present the values of the co-efficients computed from the international database. Shaded boxes represent the mean plus or minus one standard deviation of these national values. Red crosses represent the values for the national data set of the country to which the report was returned.

Substantial differences between the national and international value on one or both of these indices show that the item is behaving differently in that country. This might reflect a mistranslation or another problem specific to the national version, but if the item was misbehaving in all or nearly all countries, it might reflect a specific problem in the source item and not with the national versions.

Report 4: Comparison of national and international item difficulty parameters and thresholds

Report 4 presents the item difficulty parameters and the thresholds, in the same graphic form as Report 3. Substantial differences between the national value and the international value (*i.e.* the national value mean) might be interpreted as an item-by-country interaction. Nevertheless, appropriate estimates of the item-by-country interaction are provided in Report 5.

Figure 9.4 ■ **Example of item statistics shown in Report 4**



© OECD 2005

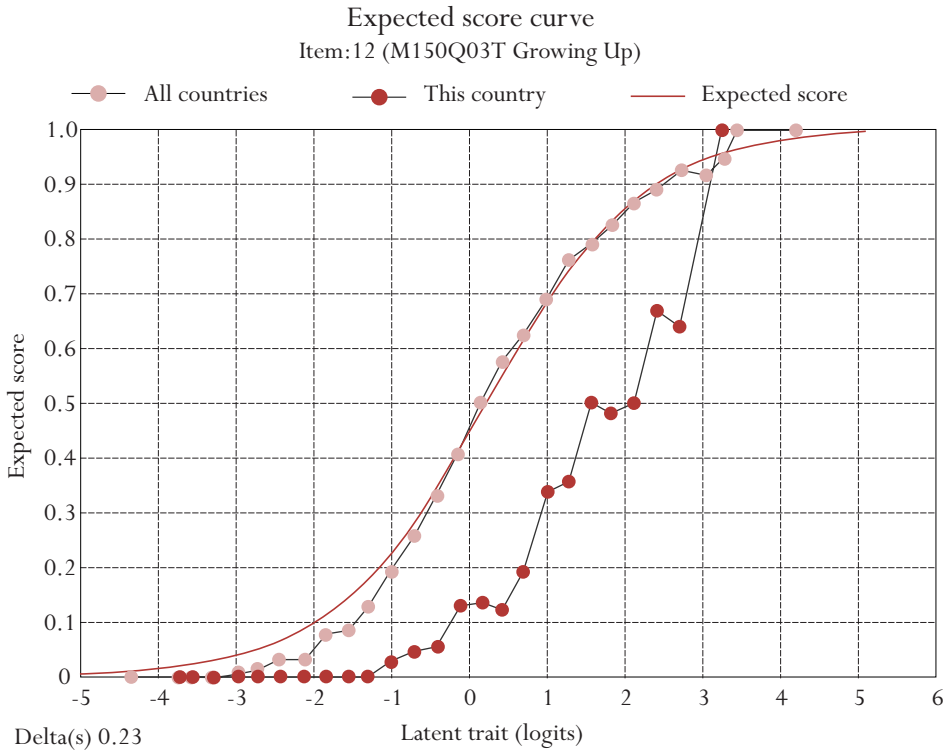
Figure 9.5 ■ **Example of item statistics shown in Report 5**

		Item by Country Interactions		Discrimination			Pisa 2000 Link Items	
	No of Valid Responses	Easier than Expected	Harder than Expected	Non-key PB is Positive	low discrimi-nation	Ability not Ordered	Link Item	Required checking
M124Q03T	1788	☐	☐	☐	☐	☐	☑	☑
M150Q03T	1601	☐	☑	☐	☐	☐	☑	☑
M155Q02T	1620	☐	☐	☐	☐	☐	☑	☑
M406Q01	1608	☑	☐	☐	☐	☐	☐	☐
M406Q02	1607	☑	☐	☐	☐	☐	☐	☐

Report 5: National dodgy item report

For each country's dodgy items, Report 5 lists where the items were flagged for one or more of the following reasons: difficulty is significantly easier or harder than average; a non-key category has a point-biserial correlation higher than 0.05 if at least 10 students selected it; the key category point-biserial correlation is lower than 0.25; the categories abilities for partial credit items are not ordered; and/or the link item difficulty was different from the PISA 2000 Main Study. An example extract is shown in Figure 9.5.

Figure 9.6 ■ **Example of item statistics shown in Report 6**



Report 6: Expected score curves

For the analysis of item performance expected score curves (ESC) were constructed and reported for each item. Report 6 provided a graphical comparison of both national and international observed scores with an expected score. Figure 9.6 is an example of the deviation of observed scores from the expected score curve. The solid line represents expected scores and the dots (connected by dotted lines) are observed scores.

International calibration

International item parameters were set by applying the conditional item response model (9.3) in conjunction with the multivariate population model (9.9), without using conditioning variables, to a sub-sample of students. This sub-sample of students, referred to as the international calibration sample, consisted of 15 000 students comprising 500 students drawn at random from each of the 30 participating OECD countries.[1]

The allocation of each PISA item to one of the seven PISA 2003 scales is given in Appendix 12 (for mathematics), Appendix 13 (for reading), Appendix 14 (for science) and Appendix 15 (for problem solving).

Student score generation

As with all item response scaling models, student proficiencies (or measures) are not observed; they are missing data that must be inferred from the observed item responses. There are several possible alternative approaches for making this inference. PISA uses the imputation methodology usually referred to as plausible values (PVs). PVs are a selection of likely proficiencies for students that attained each score.

Plausible values

Using item parameters anchored at their estimated values from the international calibration, the plausible values are random draws from the marginal posterior of the latent distribution (9.11), for each student. For details on the uses of plausible values, see Mislevy (1991) and Mislevy *et al*. (1992).

In PISA, the random draws from the marginal posterior distribution are taken as follows.

M vector-valued random deviates, $\{\varphi_{mn}\}_{m=1}^{M}$, from the multivariate normal distribution, $f_\theta(\theta_n; \mathrm{W}_n, \gamma, \Sigma)$ for each case n.[2] These vectors are used to approximate the integral in the denominator of (9.11), using the Monte-Carlo integration

$$\int_\theta f_x(x;\xi\,|\theta) f_\theta(\theta,\gamma,\Sigma) d\theta \approx \frac{1}{M}\sum_{m=1}^{M} f_x(x;\xi\,|\varphi_{mn}) \equiv \Im \qquad (9.12)$$

At the same time, the values

$$p_{mn} = f_x(x_n;\xi\,|\varphi_{mn}) f_\theta(\varphi_{mn};\mathrm{W}_n,\gamma,\Sigma) \qquad (9.13)$$

are calculated, so that the set of pairs $\left(\varphi_{mn}, p_{mn}/\Im\right)_{m=1}^{M}$, which can be used as an approximation of the posterior density (9.11) is obtained; and the probability that φ_{nj} could be drawn from this density is given by

$$q_{nj} = \frac{p_{mn}}{\sum_{m=1}^{M} p_{mn}} \qquad (9.14)$$

At this point, L uniformly distributed random numbers $\{\eta_i\}_{i=1}^{L}$ are generated; and for each random draw, the vector, φ_{ni_0}, that satisfies the condition

$$\sum_{s=1}^{i_0-1} q_{sn} < \eta_i \leq \sum_{s=1}^{i_0} q_{sn} \qquad (9.15)$$

is selected as a plausible vector.

Constructing conditioning variables

The PISA conditioning variables are prepared using procedures based on those used in the United States National Assessment of Educational Progress (NAEP) (Beaton, 1987) and in TIMSS (Macaskill, Adams and Wu, 1998). The steps involved in this process are:

- *Step 1*: Five variables (booklet ID, gender, mother's occupation, father's occupations and school mean mathematics score) were prepared to be directly used as conditioning variables. The booklet ID was dummy coded so that booklet 9 was used as the reference booklet. Booklet 9 had to be chosen as the reference booklet because it is the only booklet that contains items from all four assessment domains. For mother's and father's occupation the ISEI index was used. For each student the mean mathematics achievement for that student's school was estimated using the mean of the weighted likelihood estimates for mathematics for each of the students that also attended that student's school.
- *Step 2*: Each variable in the Student Questionnaire was dummy coded. The details of this dummy coding are provided in Appendix 10.
- *Step 3*: For each country, a principal components analysis of the dummy-coded variables was performed, and component scores were produced for each student (a sufficient number of components to account for 95 per cent of the variance in the original variables).
- *Step 4*: The item-response model was fit to each national data set and the national population parameters were estimated using item parameters anchored at their international location and conditioning variables derived from the national principal components analysis and from step 1.
- *Step 5*: Five vectors of plausible values were drawn using the method described above. The vectors were of length seven, one for each of the PISA 2003 reporting scales.

As described in Chapter 2, the PISA test design is such that not all students are assessed in all four domains. In PISA 2000, the plausible values for those students who did not respond to any items from a domain were removed from the database and a set of weight adjustments were provided for dealing with the smaller data set. The assumption under this approach is that the students who did not get domain scores were missing at random. For PISA 2003, the plausible values for all domains have been retained for all students. This approach has a number of advantages. First, the database structure is simpler and analysis is simpler because the use of a weight adjustment is not necessary. Second, the missing at random assumption is loosened somewhat. The plausible value generation assumes that the relationships between the domain for which no items are observed and all other variables (both conditioning variables and the other domain) is the same for both the students who did respond to items from a domain and those that did not. Using all of this relationship information, and all available information about the student an imputation is made. Because of the amount of data that is available to make the imputation, the analysis of the full data set will produce more accurate results than will analysis of the data set that omits students who did not respond to a domain. Additionally it can be expected that, due to sampling variation, the characteristics of the students who did not

respond to a domain will be slightly different to the characteristics of those that did. These differences will be appropriately adjusted for in the imputation and the estimated characteristics of, for example, the reading proficiency distribution for all students will be slightly different to the estimated characteristics of the reading proficiency distribution for the subset of students that responded to reading items.

The one disadvantage of this approach is that the average performances on a reference booklet (booklet 9) will influence the imputations for students who did not respond to items from a domain. As we note in Chapter 13, booklet- and item-by-country interactions do result in variations across booklets in the country means. If a country has an unusually low or high performance on the reference booklet, for a particular domain, then this unusual performance will influence the imputations for all students that did not respond to that domain. The consequential effect is that the reference booklet will be given more weight than the other booklets in the assessment of national means.

ANALYSIS OF DATA WITH PLAUSIBLE VALUES

It is important to recognise that plausible values are not test scores and should not be treated as such. They are random numbers drawn from the distribution of scores that could be reasonably assigned to each individual—that is, the marginal posterior distribution (9.11). As such, plausible values contain random error variance components and are not optimal as scores for individuals. Plausible values as a set are better suited to describing the performance of the population. This approach, developed by Mislevy and Sheehan (1987, 1989) and based on the imputation theory of Rubin (1987), produces consistent estimators of population parameters. Plausible values are intermediate values provided to obtain consistent estimates of population parameters using standard statistical analysis software such as SPSS and SAS. As an alternative, analyses can be completed using *ConQuest* (Wu *et al.*, 1997a).

The PISA student file contains 40 plausible values, five for each of the seven PISA 2003 cognitive scales and five for the combined mathematics scale. *PV1MATH* to *PV5MATH* are five for mathematical literacy; *PV1SCIE* to *PV5SCIE* for scientific literacy, *PV1READ* to *PV5READ* for reading literacy and *PV1PROB* to *PV5PROB* for problem solving. For the four mathematics literacy subscales – space and shape, change and relationship, uncertainty and quantity – the plausible values variables are *PV1MATH1* to *PV5MATH1*, *PV1MATH2* to *PV5MATH2, PV1MATH3* to *PV5MATH3,* and *PV1MATH4* to *PV5MATH4,* respectively.

If an analysis were to be undertaken with one of these seven cognitive scales, or for the combined mathematics scale, then it would ideally be undertaken five times, once with each relevant plausible values variable. The results would be averaged, and then significance tests adjusting for variation between the five sets of results computed.

More formally, suppose that $r(\theta, \mathrm{Y})$ is a statistic that depends upon the latent variable and some other observed characteristic of each student. That is: $(\theta, \mathrm{Y}) = (\theta_1, y_1, \theta_2, y_2, \ldots, \theta_N, y_N)$ where (θ_n, y_n) are the values of the latent variable and the other observed characteristic for student n. Unfortunately θ_n is not observed, although we do observe the item responses, x_n from which we can construct for each student n, the marginal posterior $h_\theta(\theta_n; y_n, \xi, \gamma, \Sigma \,|\mathrm{x}_n)$. If $h_\theta(\theta; \mathrm{Y}, \xi, \gamma, \Sigma \,|\mathrm{X})$ is the joint marginal posterior for $n=1,\ldots N$ then we can compute:

$$r^*(\mathrm{X}, \mathrm{Y}) = E\left[r^*(\theta, \mathrm{Y}) \middle| \mathrm{X}, \mathrm{Y}\right] = \int_\theta r(\theta, \mathrm{Y}) h_\theta(\theta; \mathrm{Y}, \xi, \gamma, \Sigma \,|\mathrm{X}) d\theta \qquad (9.16)$$

The integral (9.16) can be computed using the Monte-Carlo method. If M random vectors $(\Theta_1, \Theta_2, \ldots, \Theta_M)$ are drawn from $h_\theta(\theta; Y, \xi, \gamma, \Sigma | X)$ (9.16) is approximated by:

$$r^*(X,Y) \approx \frac{1}{M}\sum_{m=1}^{M} r(\Theta_m, Y) = \frac{1}{M}\sum_{m=1}^{M} \hat{r}_m \qquad (9.17)$$

where $\hat{r}_m$ is the estimate of r computed using the m-th set of plausible values.

From (9.16) we can see that the final estimate of r is the average of the estimates computed using each plausible value in turn. If U_m is the sampling variance for $\hat{r}_m$ then the sampling variance of r^* is:

$$V = U^* + (1 + M^{-1})B_M \qquad (9.18)$$

where $U^* = \frac{1}{M}\sum_{m=1}^{M} U_m$ and $B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{r}_m - r^*)^2$.

An α-% confidence interval for r^* is $r^* \pm t_\upsilon\left(\frac{(1-\alpha)}{2}\right)V^{1/2}$

where $t_\upsilon(s)$ is the s-percentile of the t-distribution with V degrees of freedom. $\upsilon = \left[\frac{f_M^2}{M-1} + \frac{(1-f_M)^2}{d}\right]^{-1}$, $f_M = (1 + M^{-1})B_M / V$ and d is the degree of freedom that would have applied had θ_n been observed. In PISA, d will vary by country and have a maximum possible value of 80.

DEVELOPING COMMON SCALES FOR THE PURPOSES OF TRENDS

The reporting scales that were developed for each of reading, mathematics and science in PISA 2000 were linear transformations of the natural logit metrics that result from the scaling as described above. The transformations were chosen so that the mean and standard deviation of the PISA 2000 scores was 500 and 100 respectively, for the 27 OECD countries that participated in PISA 2000 that had acceptable response rates (see Adams and Wu, 2002).[3]

For PISA 2003, the decision was made to report the reading and science scores on these previously developed scales. That is the reading and science reporting scales used for PISA 2000 and PISA 2003 are directly comparable. The value of 500, for example, has the same meaning as it did in PISA 2000 – that is, the mean score in 2000 of the sampled students in the 27 OECD countries that participated in PISA 2000.[4]

For problem solving, which is a new domain for PISA 2003, and for mathematics this is not the case, however. Mathematics, as the major domain, was the subject of major development work for PISA 2003, and the PISA 2003 mathematics assessment was much more comprehensive than the PISA 2000 mathematics assessment – the PISA 2000 assessment covered just two (space and shape, and change and relationships) of the four areas that are covered in PISA 2003. Because of this broadening in the assessment it was deemed inappropriate to report the PISA 2003 mathematics scores on the same scale as the PISA 2000 mathematics scores. For both problem solving and mathematics the linear transformation of the logit metric was chosen such that the mean was 500 and standard deviation 100 for the 30 OECD countries that participated in PISA 2003.[5]

Linking PISA 2000 and PISA 2003 reading and science

The PISA 2000 and PISA 2003 assessments of mathematics, reading and science are linked assessments. That is, the sets of items used to assess each of mathematics, reading and science in PISA 2000 and the sets of items used to assess each of mathematics, reading and science in PISA 2003 include a subset of items common to both sets. For mathematics 20 items were used in both assessments, for reading 28 items were used in both assessments and for science 25 items were used in both assessments (see Chapter 2). These common items are referred to as link items.

The steps involved in linking the PISA 2000 and PISA 2003 reading and science scales were:

- *Step 1*: The PISA 2000 data from each of the OECD countries were then re-scaled with full conditioning and with link items anchored at their PISA 2003 values.
- *Step 2*: The mean and standard deviation of each domain were calculated for a combined data set of 25 OECD countries[6]. Senate weights were used so that each country was given the same weight.
- *Step 3*: The mean and standard deviations computed in Step 2 were then compared with the matching means and standard deviations from the PISA 2000 scaling. Linear transformations that mapped the PISA 2003 based scores to scores that would yield a mean and standard deviation equal to the PISA 2000 results were then computed.

Linking PISA 2000 and PISA 2003 mathematics

In the case of mathematics a decision was made to produce a new scale for PISA 2003 and to undertake a retrospective mapping of the 2000 data onto this new PISA 2003 scale for each of the two areas (space and shape, and change and relationships) that were assessed both times. The steps involved were:

- *Step 1*: The PISA 2000 calibration sample was scaled with a two dimensional model, the two dimensions being the two mathematics scales included in PISA 2000. The items were anchored at their PISA 2000 values. No conditioning was used in this scaling.
- *Step 2*: Step 1 was then replicated with the items anchored at their PISA 2003 values.
- *Step 3*: For the two sets of scaling results the means and standard deviations for both dimensions were calculated for a combined data set of 25 OECD countries.[7] Senate weights were used so that each country was given the same weight.
- *Step 4*: Linear transformations that mapped the PISA 2000 based scores to scores that would yield a mean and standard deviation equal to the PISA 2003 results were then computed.

Uncertainty in the link

In each case the transformation that equates the 2000 and 2003 data depends upon the change in difficulty of each of the individual link items and as a consequence the sample of link items that has been chosen will influence the choice of transformation. This means that if an alternative set of link items had been chosen the resulting transformation would be slightly different. The consequence is an uncertainty in the transformation due to the sampling of the link items, just as there is an uncertainty in values such as country means due to the use of a sample of students.

The uncertainty that results from the link-item sampling is referred to as linking error and this error must be taken into account when making certain comparisons between PISA 2000 and PISA 2003 results. Just as with the error that is introduced through the process of sampling students, the exact magnitude of this linking error cannot be determined. The likely range of magnitudes for this error can, however, be estimated and this error can be taken into account when interpreting PISA results. As with sampling errors, the likely range of magnitude for the errors is represented as a standard error. The link standard errors are reported in Chapter 13.

In PISA a common transformation has been estimated, from the link items, and this transformation is applied to all participating countries. It follows that any uncertainty that is introduced through the linking is common to all students and all countries. Thus, for example, suppose the unknown linking error (between PISA 2000 and PISA 2003) in reading resulted in an over-estimation of student scores by two points on the PISA 2000 scale. It follows that every student's score will be over-estimated by two score points. This over-estimation will have effects on certain, but not all, summary statistics computed from the PISA 2003 data. For example, consider the following:

- Each country's mean will be over-estimated by an amount equal to the link error. In this example, it is two score points.
- The mean performance of any subgroup will be over-estimated by an amount equal to the link error. In this example, it is two score points.
- The standard deviation of student scores will not be affected because the over-estimation of each student by a common error does not change the standard deviation.
- The difference between the mean scores of two countries in PISA 2003 will not be influenced because the over-estimation of each student by a common error will have distorted each country's mean by the same amount.
- The difference between the mean scores of two groups (*e.g.* males and females) in PISA 2003 will not be influenced, because the over-estimation of each student by a common error will have distorted each group's mean by the same amount
- The difference between the performance of a group of students (*e.g.* a country) between PISA 2000 and PISA 2003 will be influenced because each student's score in PISA 2003 will be influenced by the error.
- A change in the difference in performance between two groups from PISA 2000 to PISA 2003 will not be influenced. This is because neither of the components of this comparison, which are differences in scores in 2000 and 2003 respectively, is influenced by a common error that is added to all student scores in PISA 2003.

In general terms, the linking error need only be considered when comparisons are being made between PISA 2000 and PISA 2003 results, and then usually only when group means are being compared.

The most obvious example of a situation where there is a need to use linking error is in the comparison of the mean performance for a country between PISA 2000 and PISA 2003. For example, let us consider a comparison between 2000 and 2003 of the performance of Denmark in reading. The mean performance of Denmark in 2000 was 497 with a standard error of 2.4, while in 2003 the mean was 492 with a standard error of 2.8. The standardised difference in the mean for Denmark is 0.89, which is computed as follows: $0.89 = (497 - 492) / \sqrt{2.4^2 + 2.8^2 + 3.744^2}$, and is not statistically significant.

Notes

1 The samples used were simple random samples stratified by the explicit strata used in each country. Students who responded to the UH booklet were not included in this process.

2 The value M should be large. For PISA, 2000 has been used.

3 Using senate weights.

4 Using senate weights.

5 Using senate weights.

6 The Netherlands was excluded because it did not meet PISA standards in 2000. The United Kingdom was excluded because it did not meet PISA standards in 2003. Luxembourg was omitted because of a change in test administration procedures between PISA 2000 and 2003. The Slovak Republic and Turkey were excluded because they did not participate in PISA 2000.

7 See footnote 6.

10

Coding Reliability Studies

As described in Chapter 2, a substantial proportion of the PISA 2003 items were open ended and required coding by trained personnel. It was important therefore that PISA implemented procedures that maximised the validity and consistency, both within and between countries, of this coding. Each country coded items on the basis of coding guides prepared by the consortium (see Chapter 2) using the coding design described in Chapter 6. Training sessions to train countries in the use of the coding guides were held prior to both the field trial and the main study.

This chapter describes three aspects of the coding and coding reliability studies undertaken in conjunction with the field trial and the main study. These are:

- The homogeneity analyses undertaken with the field trial data to assist the test developers in constructing valid, reliable scoring rubrics;
- The variance component analyses undertaken with the main study data to examine within-country rater reliability; and,
- An inter-country reliability study undertaken to examine the between-country consistency in applying the coding guides.

EXAMINING WITHIN-COUNTRY VARIABILITY IN CODING

To obtain an estimate of the between-marker variability within each country, multiple coding was required for at least some student answers. Therefore, it was decided that multiple codings would be collected for all open-ended items in both the field trial and the main study for a moderate number of students. In the main study, 100 students' booklets were multiple coded. The requirement was that the same four expert markers per domain (reading, mathematics and science) should mark all items appearing together in the first two clusters of the test booklets 1 to 6, 8, 10 and 12. A booklet 10 containing, for example, 14 reading items, would give a three-dimensional table for reading (100 students by 14 items by 4 coders), where each cell contains a single category. For each domain and each booklet, such a table was produced for each country. The following describes the various analyses of these data.

The field trial problems were quite different from those in the main study. In the field trial, many more items were tried than were used in the main study. One important purpose of the field trial was to select a subset of items to be used in the main study. One obvious concern was to ensure that markers agreed, to a reasonable degree, in their categorisation of the answers. More subtle problems can arise, however. In the final administration of a test, a student answer is scored numerically. But in the construction phase of an item, more than two response categories may be provided, say A, B and C, and it may not always be clear how these should be converted to numerical scores. The technique used to analyse the field trial data can provide at least a partial answer, and also give an indication of the agreement between markers for each item separately. The technique is called homogeneity analysis.

In the main study, the problem was different. The field trial concluded with a selection of a definite subset of items and a scoring rule for each. The major problem in the main study was to determine how much of the total variance of the numerical test scores could be attributed to variability across coders. The basic data set to be analysed therefore consisted of a three dimensional table of numerical scores. The technique used is referred to as variance component analysis.

Some items in the field trial appeared to function poorly because of well-identified defects, *e.g.* poor translation). To get an impression of the differences in marker variability between field trial and main study,

most field trial data analyses were repeated using the main study data. Some comparisons are reported below. This chapter uses a consistent notational system, summarised in Figure 10.1. Nested data structures are occasionally referred to, as every student and every marker belong to a single country. In such cases, the indices m and v take the subscript c.

Figure 10.1 ■ **Notation system**

Symbol	Range	Meaning
i	$1,\ldots,I$	Item
c	$1,\ldots,C$	Country
V_c		Number of students from country c
v	$1,\ldots,V_c$	Student
M_c		Number of coders from country c
m	$1,\ldots,M = \sum_c M_c$	Coder
l	$1,\ldots,L_i$	Category of item i

Homogeneity analysis

In the analysis, the basic observation is the category into which coder m places the response of student v on item i, denoted O_{ivm}. Basic in the approach of homogeneity analysis is to consider observations as qualitative or nominal variables. (For a more mathematical treatment of homogeneity analysis, see Nishisato, 1980; Gifi, 1990; or Greenacre, 1984.) Although observations may be coded as digits, these digits are considered as labels, not numbers. To have a consistent notational system, it is assumed in the sequel that the response categories of item i are labelled $1\ldots l\ldots L_i$. The main purpose of the analysis is to convert these qualitative observations into (quantitative) data which are in some sense optimal.

The basic loss function

The first step in the analysis is to define a set of (binary) indicator variables that contain all the information of the original observations, defined by:

$$O_{ivm} = l \Leftrightarrow g_{ivml} = 1 \qquad (10.1)$$

where it is to be understood that g_{ivml} can take only the values 0 and 1.

The basic principle of homogeneity analysis is to assign a number x_{iv} to each student (the student score on item i), and a number y_{iml} to each observation O_{ivm}, called the category quantification, such that student scores are in some way the best summary of all category quantifications that apply to them. To understand this in more detail, consider the following loss function:

$$F_i = \sum_v \sum_m \sum_l g_{ivml} \left(x_{iv} - y_{iml}\right)^2 \qquad (10.2)$$

The data are represented by indicator variables g_{ivml}. If coder m has a good idea of the potential of student v, and thinks it appropriate to assign him or her to category l, then, ideally, one would expect that $x_{iv} = y_{iml}$, yielding a zero loss for that case. But the same coder can have used the same category for student v', who has a different potential $x_{iv'}$, and since the quantification y_{iml} is unique, there cannot be a zero loss in both cases. Thus, some kind of compromise is required, which is made by minimising the loss function F_i.

Four observations need to be made in connection with this minimisation:

- The loss function (10.2) certainly has no minimum, because adding an arbitrary constant to all x_{iv} and all y_{iml} leaves F_i unaltered. This means that in some way an origin of the scale must be chosen. Although this origin is arbitrary, there are some theoretical advantages in defining it through the equality:

$$\sum_v x_{iv} = 0 \qquad (10.3)$$

- If for all v and all m and one chooses $x_{iv} = y_{iml} = 0$ then $F_i = 0$ (and 10.3 is fulfilled), which is a minimum. Such a solution, where all variability in the student scores and in the category quantifications is suppressed, is called a degenerate solution. To avoid such degeneracy, and at the same time to choose a unit of the scale, requires the restriction:

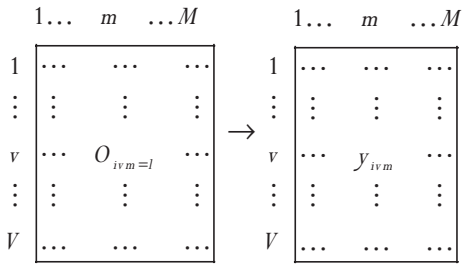
$$\frac{1}{V}\sum_v x_{iv}^2 = 1 \qquad (10.4)$$

Restrictions (10.3) and (10.4) jointly guarantee that a unique minimum of F_i exists and corresponds to a non-degenerate solution except in some special cases, as discussed below.

- Notice that in the loss function, missing observations are taken into account in an appropriate way. From definition (10.1) it follows that if O_{ivm} is missing, $g_{viml} = 0$ for all l, such that a missing observation never contributes to a positive loss.
- A distinct loss function is minimised for each item. Although other approaches to homogeneity analysis are possible, the present one serves the purpose of item analysis well. The data pertaining to a single item are analysed separately, requiring no assumptions on their relationships. A later subsection shows how to combine these separate analyses to compare markers and countries.

Another way to look at homogeneity analysis is to arrange the basic observations (for a single item i) in a table with rows corresponding to students and columns corresponding to markers, as in the left panel of Figure 10.2. The results can be considered as a transformation of the observations into numbers, as shown in the right panel of the figure. At the minimum of the loss function, the quantified observations y_{iml} have the following properties.

Figure 10.2 ■ **Quantification of categories**



© OECD 2005

The total variance can be partitioned into three parts: one part attributable to the columns (the markers), another part attributable to the rows (the students), and a residual variance. At the solution point it holds that:

$$\begin{aligned} &Var\left(students\right) \text{is maximised} \\ &Var\left(coders\right)=0 \\ &Var\left(residuals\right) \text{is minimised} \end{aligned} \qquad (10.5)$$

If the coders have a high agreement among themselves, *Var(residuals)* will be a small proportion of the total variance, meaning that the markers are very homogeneous. The index of homogeneity is defined therefore as:

$$H_{ic} = \frac{Var\left(students\right)}{Var\left(students\right)+Var\left(residuals\right)} \qquad (10.6)$$

at the point where F_i attains its minimum. The subscript c has been added to indicate that this index of homogeneity can only be meaningfully computed within a single country.

The indices H_{ic} can be compared meaningfully with each other, and across countries and items, because they are all proportions of variance attributable to the same source (students), compared to the total variance attributable to students and markers. The differences between the H_{ic}-indices, therefore, must be attributed to the items, and can therefore be used as an instrument for item analysis. Items with a high H_{ic} index are less susceptible to marker variation, and therefore the scores obtained on them are more easily generalisable across markers.

Degenerate and quasi-degenerate solutions

Although restriction (10.4) was introduced to avoid degenerate solutions, it is not always sufficient, and the data collection design or some peculiarities in the collected data can lead to other kinds of degeneration. First, degeneracy due to the design is discussed.

Using the notational convention explained in the introduction, the loss function (10.2) can, in the case of the data of all countries analysed jointly, be written as:

$$F_i = \sum_{c}\sum_{v_c}\sum_{m_c}\sum_{l} g_{iv_c m_c l}\left(x_{iv_c} - y_{im_c l}\right)^2 \qquad (10.7)$$

To see the degeneracy clearly, suppose $C = 2$, $V_1 = V_2$ and there are no missing responses. The value F_i=0 (and thus $H_i = 1$) can be attained as follows: $x_{iv_c} = 1$, $c = 1$ if ,$x_{iv_c} = -1$, if $c = 2$ and $y_{im_c l} = x_{iv_c}$ (all l). This solution complies with (10.3) and (10.4), but it can easily be seen that it does nothing other than maximise the variance of the x's between countries and minimise the variance within countries. Of course, one could impose a restriction analogous to (10.4) for each country, but then (10.7) becomes C independent sums, and the scores (x-values) are no longer comparable across countries. A meaningful comparison across countries requires imposing restrictions of another kind, such restrictions are described below.

In some cases, this kind of degeneracy may occur also in data sets collected in a complete design, but with

extreme patterns of missing observations. Suppose some student has got a code from only one marker and assume, moreover, that this marker used this code only once. A degenerate solution may then occur where this student is contrasted with all others (collapsed into a single point), much in the same way as in the example above.

Similar cases may occur when a certain code, l say, is used a very few times. Assume this code is used only once, by marker m. By choosing a very extreme value for y_{im_cl}, a situation may occur where the student who is given code by marker m tends to contrast with all the others, although fully collapsing them may be avoided (because this student's score is pulled towards the others by the codes received from the other markers). But the general result will be one where the solution is dominated by this very infrequent coding by one marker. Such cases may be called quasi-degenerate and are examples of chance capitalisation. They are prone to occur in small samples of students, especially in cases where there are many different categories – as in the field trial, especially with items with multiple-answer categories. Cases of quasi-degeneracy give a spuriously high H_i index, and as such the high index needs to be interpreted with caution.

Quasi-degeneracy is an intuitive notion that is not rigorously defined, and will continue to be a major source of concern, although adequate restrictions on the model parameters usually address the problem.

To develop guidelines for selecting a good test from the many items used in the field trial, one should realise that a low homogeneity index points to items that will introduce considerable variability into the test score because of rater variance, and may therefore best be excluded from the definitive test. But an item with a high index is not necessarily a good item. Quasi-degeneracy tends to occur in cases where one or more response categories are used very infrequently. It might therefore be useful to develop a device that can simultaneously judge homogeneity and the risk of quasi-degeneracy.

Homogeneity analysis with restrictions

Apart from cases of quasi-degeneracy, there is another reason for imposing restrictions on the model parameters in homogeneity analysis. The specific value of H_i, obtained from the minimisation of (10.2), can only be attained if the quantification issued from the homogeneity analysis is indeed used in applications, *i.e.* when the score obtained by student v on item i when categorised by marker m in category l is equal to the category quantification y_{im_cl}. But this means that the number of points to be earned from receiving category l may differ across markers. An extra difficulty arises when new markers are used in future applications. This would imply that in every application a new homogeneity analysis has to be completed. And this is not very attractive for a project like PISA, where the field trial is meant to determine a definite scoring rule, whereby the same scoring applies across all coders and countries.

Restrictions within countries

As a first restriction, one might wish for no variation across markers within the same country so that for each country c the restriction:

$$y_{im_cl} = y_{icl}, \left(m = 1,\ldots,M_c; l = 1,\ldots,L_i\right) \qquad (10.8)$$

is imposed. But since students and markers are nested within countries, this amounts to minimising the loss function for each country:

$$F_{ic}^{*} = \sum_{v_c}\sum_{m_c}\sum_{l} g_{iv_cm_cl}\left(x_{iv_c}^{*} - y_{icl}\right)^2 \qquad (10.9)$$

© OECD 2005

such that the overall loss function is minimised automatically to:

$$F_i^* = \sum_c F_{ic}^* \qquad (10.10)$$

To minimise (10.9), technical restrictions similar to (10.3) and (10.4) must hold per country. As in the case without restrictions, homogeneity indices can be computed in this case also (see equation (10.6)), and will be denoted . It is easy to understand that for all items and all countries the inequality:

$$H_{ic}^* \le H_{ic} \qquad (10.11)$$

must hold.

In contrast to some indices to be discussed further, H_{ic}^* indices are not systematically influenced by the numbers of students or coders. If students and coders within a country can be considered to be a random sample of some populations of students and markers, the computed indices are a consistent (and unbiased) estimate of their population counterparts. The number of students and coders influence only the standard error.

The H_{ic}^* indices can be used for a double purpose:

- Comparing H_{ic}^* with H_{ic} within countries: if for some item H_{ic}^*, is much lower than H_{ic}, this may point to systematic differences in interpretation of the coding guides among markers. Suppose, as an example, a binary item with categories A and B, and all markers in a country except one agree perfectly among themselves in their assignment of students to these categories; one marker, however, disagrees perfectly with the others in assigning category A where the others choose *B*, and *vice versa*. Since each marker partitions all students in the same two subsets, the index of homogeneity H_{ic} for this item will be one, but the category quantifications of the outlying marker will be different from those of the other markers. Requiring that the category quantifications be the same for all markers will force some compromise, and the resulting H_{ic}^* will be lower than one.
- Differences between H_{ic}^* and H_{ic} can also be used to compare different countries among each other (per item). Differences, especially if they are persistent across items, make it possible to detect countries where the coding process reveals systematic disagreement among the coders.

Restrictions within and across countries

Of course, different scoring rules for different countries are not acceptable for the main study. More restrictions are therefore needed to ascertain that the same category should correspond to the same item score (quantification) in each country. This amounts to a further restriction on (10.8), namely:

$$y_{icl} = y_{il}, (c = 1,\ldots,C; l = 1,\ldots,L_i) \qquad (10.12)$$

leading to the loss function:

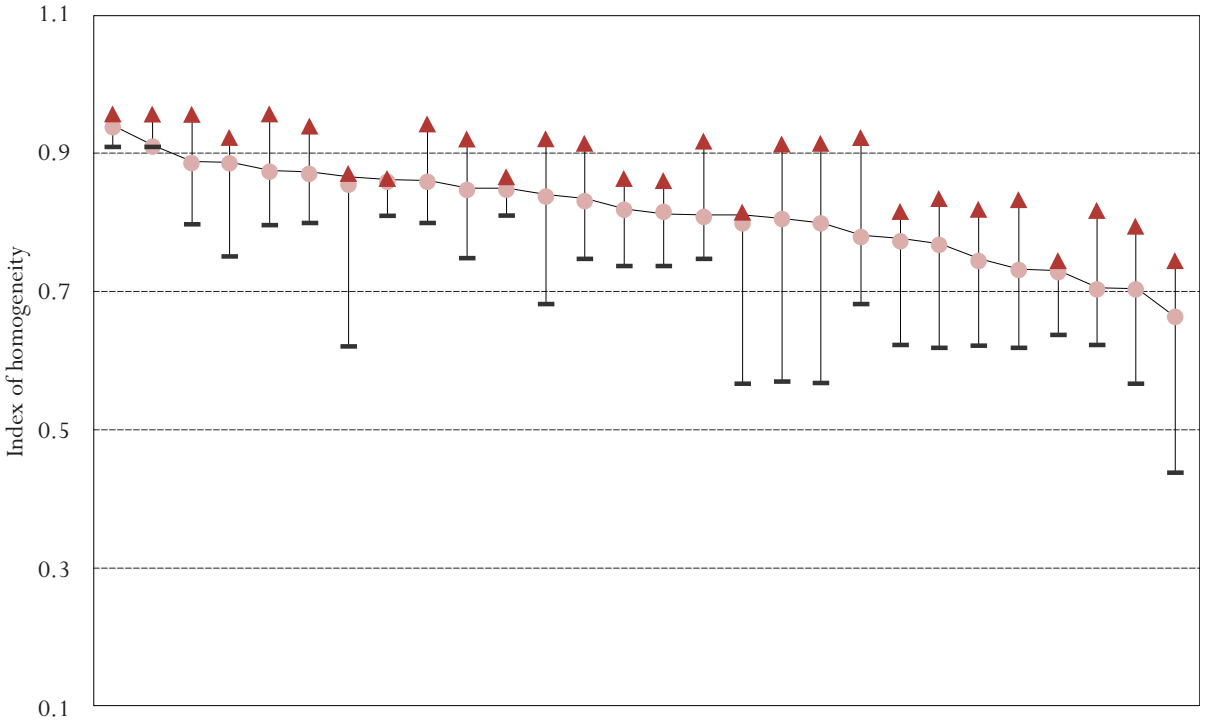
$$F_i^{**} = \sum_c \sum_{v_c} \sum_{m_c} \sum_l g_{iv_c m_c l} \left(x_{iv_c}^{**} - y_{il}\right)^2 \qquad (10.13)$$

and a corresponding index of homogeneity, denoted as H_i^{**}.

To provide an impression of the use of these indices, a summary plot for the open-ended field trial items in science is displayed in Figure 10.3. In the field trial each country was allocated into one of two groups *A*

and *B*. The countries in group *A* have multiple marked the first two clusters from all of the odd-numbered booklets to be marked. The countries in group *B* multiple-marked the first two clusters from all of the even-numbered booklets to be marked. The line shown in the chart connects the H_{ic}^{**} indices for all items (sorted in decreasing order). An interval is plotted, for each item, based on the H_{ic}^{*} indices of 23 or 21 countries depending on group they were allocated. For these countries, the H_{ic}^{*} indices were sorted in ascending order and the endpoints of the intervals correspond to the inter-quartile range of the indices. This figure was a useful guide for selecting items for the main study since it allowed a selection based on the value of H_{ic}^{**}, but also shows clear differences in the variability of the H_{ic}^{*} indices. The first three items, for example, have almost equal H_{ic}^{**} indices, but the third one shows more variability across countries than the other two, and therefore may be less suitable for the main study. But perhaps the clearest example is the last one, with the lowest H_{ic}^{**} index and showing large variations across countries.

Figure 10.3 ■ H_{ic}^{*} **and** H_{ic}^{**} **for science items in the field trial**



An additional criterion for selecting items

An ideal situation (from the viewpoint of statistical stability of the estimates) occurs when each of the L_i response categories of an item has been used an equal number of times:

- For each marker separately when one does the analysis without restrictions;
- Across markers within a country when an analysis is done with restrictions within a country (restriction (10.8)); or
- Across markers and countries when restriction (10.12) applies.

If the distribution of the categories departs strongly from uniformity, cases of quasi-degeneracy may occur: the contrast between a single category with very small frequency and all the other categories may tend

to dominate the solution. Very small frequencies are more likely to occur in small samples than in large samples, hence the greater possibility of being influenced by chance. The most extreme case occurs when the whole distribution is concentrated in a single category; a homogeneity analysis thus has no meaning (and technically cannot be carried out because normalisation is not defined).

From a practical point of view, an item with a distribution very close to the extreme case of no variability is of little use in testing, but may have a very acceptable index of homogeneity. Therefore, it seems wise to judge the quality of an item by considering simultaneously the homogeneity index and the form of distribution. To measure the departure from a uniform distribution in the case of nominal variables, the index to be developed must be invariant under permutation of the categories. For a binary item, for example, the index for an item with p-value p must be equal to that of an item with p-value $1-p$.

Pearson's well-known X^2 statistic, computed with all expected frequencies equal to each other, fulfils this requirement, and can be written in formula form as:

$$X^2 = \sum_l \frac{\left(f_l - \bar{f}\right)^2}{\bar{f}} = n \sum_l \frac{\left(p_l - \bar{p}\right)^2}{\bar{p}} \qquad (10.14)$$

where, in the middle expression, f_l is the frequency of category l, $\bar{f}$ is the average frequency and, in the right-hand expression p_l, and $\bar{p}$ are observed and average proportions, and n is the sample size. This index, however, changes with changing sample size, and comparison across items (even with constant sample size) is difficult because the maximum value increases with the number of categories. It can be shown that:

$$X^2 \leq n \ (L-1) \qquad (10.15)$$

where equality is reached only where $L-1$ categories have frequency zero, *i.e.* the case of no variability. The index:

$$\Delta = \frac{X^2}{n\left(L-1\right)} \qquad (10.16)$$

is proposed as an index of departure from uniformity. Its minimum is zero (uniform distribution), its maximum is one (no variability).

It is invariant under permutation of the categories and is independent of the sample size. This means that it does not change when all frequencies are multiplied by a positive constant. Using proportions p_l instead of frequencies f_l, (10.16) can be written as:

$$\Delta = \frac{L}{L-1} \sum_l \left(p_l - \frac{1}{L} \right)^2 \qquad (10.17)$$

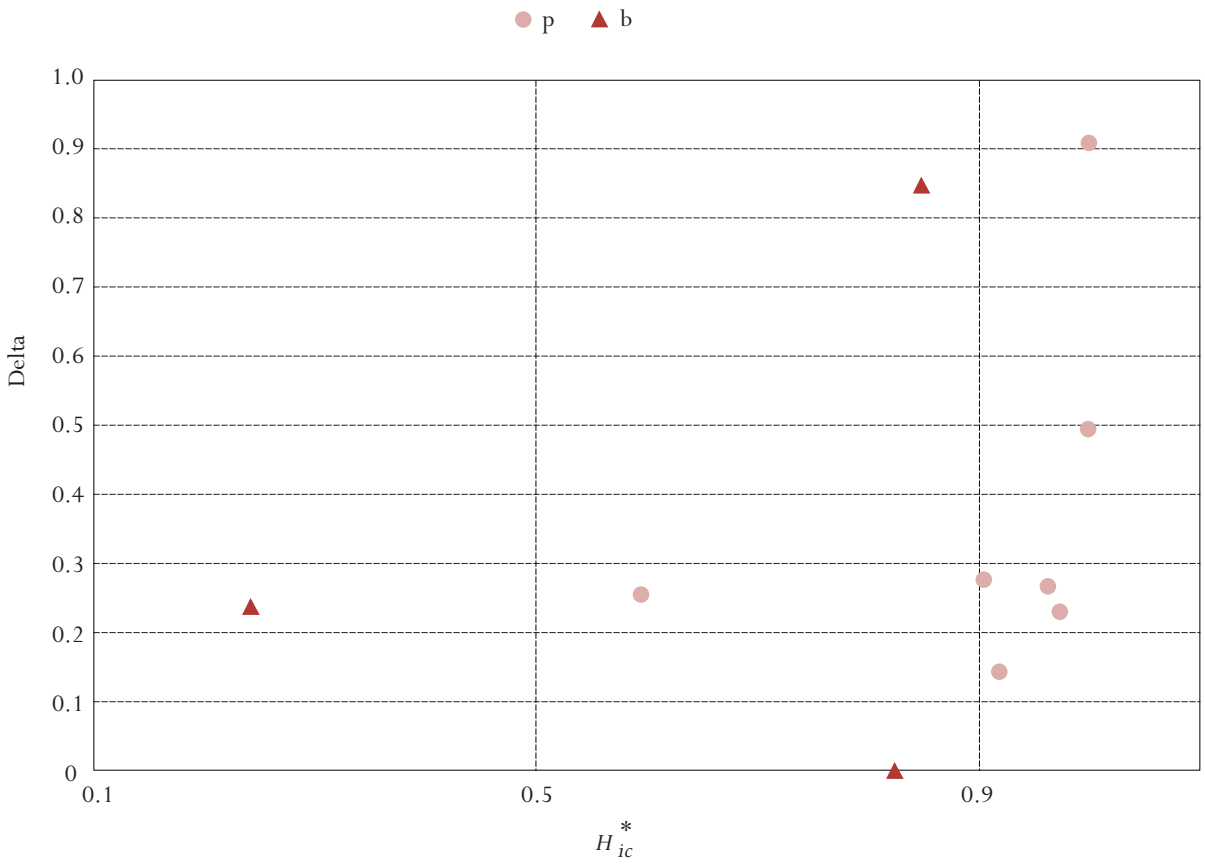
Table 10.1 gives some distributions and their associated values for $L = 2$ and $L = 3$. (Row frequencies always sum to 100.)

Table 10.1 ■ **Some examples of Δ**

L = 2			L = 3			
Frequency		Δ	Frequency			Δ
50	50	0.00	50	50	0	0.25
60	40	0.04	40	30	30	0.01
70	30	0.16	50	25	25	0.06
75	25	0.25	50	48	2	0.22
80	20	0.36	70	15	15	0.30
90	10	0.64	70	28	2	0.35
95	5	0.81	80	10	10	0.49
99	1	0.96	80	8	2	0.51

As an example, the H^*_{ic} indices are plotted in Figure 10.4 against the Δ-values for the 10 open-ended mathematics items in Booklet 3 of the field trial, using the Ireland data. The binary (b) items are distinguished from the items with more than two categories (p). One can see that 6 of the 10 items are situated in the lower right-hand corner of the figure, combining high homogeneity with a response distribution that does not deviate too far from a uniform distribution. But two items have high homogeneity indices and a very skewed distribution, which made them less suitable for inclusion in the main study.

Figure 10.4 ■ **Homogeneity and departure from uniformity**



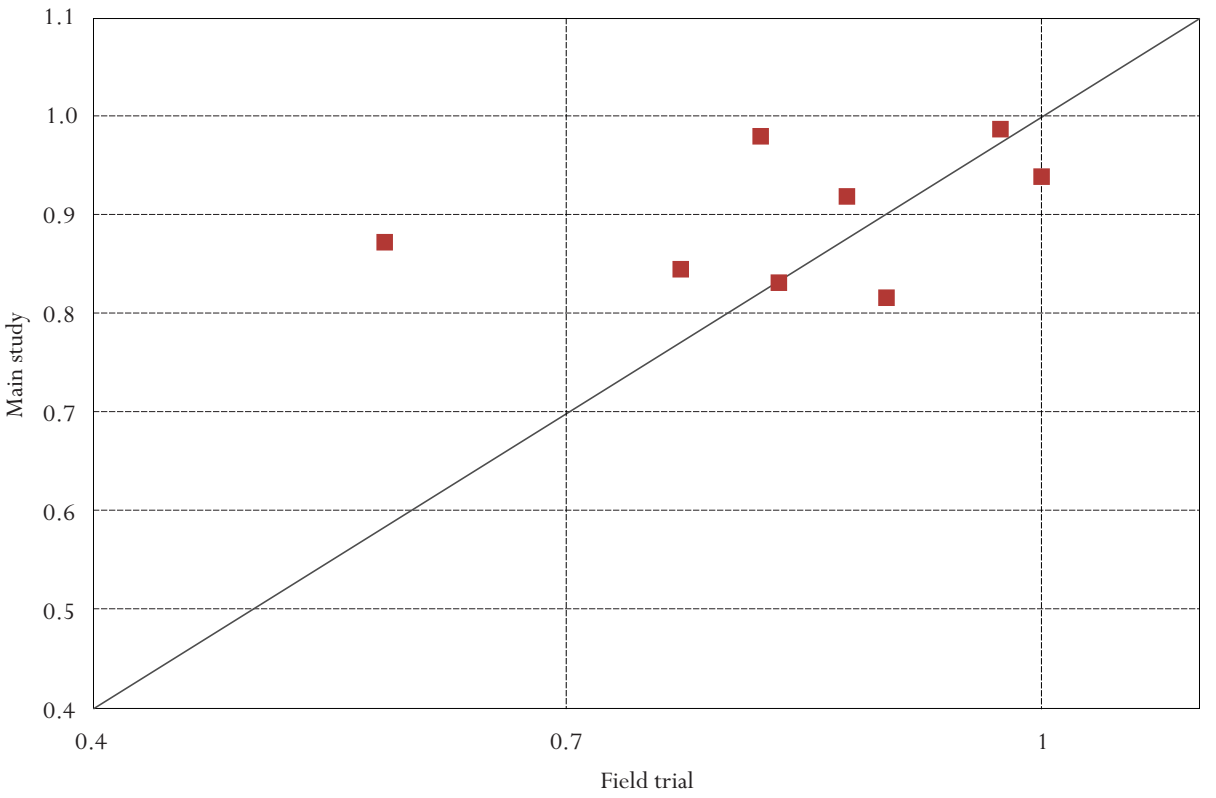
© OECD 2005

A comparison between the field trial and the main study

The main study used a selection of items from the field trial, but the coding guides was changed in some cases (mainly due to incorrect translations). The changed items were not tested in an independent field trial, however. If they did really represent an improvement, their homogeneity indices should rise in comparison with the field trial. The second reason for repeating the homogeneity analysis in the main study is the relative numbers of students and markers used for the reliability study in the field trial and in the main study. Small numbers easily give rise to chance capitalisation, and therefore repeating the homogeneity analyses in the main study serves as a cross-validation. The third reason is that PISA 2000 link items were not multiple marked during the PISA 2003 field trial and analysis had to be repeated for these items.

Figure 10.5 shows a scatter-plot of the H_{ic}^{*} indices of the eight mathematics items in the Ireland field trial and main study samples. In this example H_{ic}^{*} indices for four of items have increased in the main study.

Figure 10.5 ■ **Comparison of homogeneity indices for mathematics items in the main study and the field trial in Ireland**



Variance component analysis

The general approach to estimating the variability in the scores due to markers is generalisability theory. Introductions to the approach can be found in Cronbach *et al.* (1972), Brennan (1992), and OECD (2004). The present section, introduces the general theory and the common estimation methods. A generalisability coefficient is then derived, as a special correlation coefficient, and its interpretation is discussed. Finally some special PISA-related estimation problems are discussed.

Analysis of two-way tables: The student by items design

To make the notion of generalisability theory clear, a simple case where a number of students answer a number of items, and for each answer they get a numerical score, which will be denoted as Y_{vi}, the subscript v referring to the student, and the subscript i referring to the item, is described first. These observations can be arranged in a VxI rectangular table or matrix, and the main purpose of the analysis is to explain the variability in the table. Conceptually, the model used to explain the variability in the table is the following:

$$Y_{vi} = \mu + \alpha_v + \beta_i + (\alpha\beta)_{vi} + \varepsilon^*_{vi} \qquad (10.18)$$

where μ is an unknown constant, α_v is the student effect, β_i is the item effect, $(\alpha\beta)_{vi}$ (to be read as a single symbol, not as a product) is the student-item interaction effect, and ε^*_{vi} is the measurement error. The general approach in generalisability theory is to assume that the students in the sample are randomly drawn from some population, but also that the items are randomly drawn from a population, usually called a universe. This means that the specific value α_v can be considered as a realisation of a random variable, α for example, and similarly for the item effect: β_i is a realisation of the random variable β. Also the interaction effect $(\alpha\beta)_{vi}$ and the measurement error ε^*_{vi} are considered as realisations of random variables. So, the model says that the observed score is a sum of an unknown constant μ and four random variables.

The model as given in (10.18), however, is not sufficient to work with, for several reasons. First, since each student in the sample gives only a single response to each item in the test, the interaction effects and the measurement error are confounded. This means that there is no possibility to disentangle interaction and measurement error. Therefore, they will be taken together as a single random variable, ε for example, which is called the residual (and which is not the same as the measurement error). This is defined as:

$$\varepsilon_{vi} = (\alpha\beta)_{vi} + \varepsilon^*_{vi} \qquad (10.19)$$

and (10.18) can then be rewritten as:

$$Y_{vi} = \mu + \alpha_v + \beta_i + \varepsilon_{vi} \qquad (10.20)$$

Second, since the right-hand side of (10.20) is a sum of four terms, and only this sum is observed, the terms themselves are not identified, and therefore three identification restrictions have to be imposed. Suitable restrictions are:

$$E(\alpha) = E(\beta) = E(\varepsilon) = 0 \qquad (10.21)$$

Third, apart from the preceding restriction, which is of a technical nature, there is one important theoretical assumption: all the nonobserved random variables (student effects, item effects and residuals) are mutually independent.

This assumption leads directly to a simple variance decomposition:

$$\sigma^2_Y = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\varepsilon \qquad (10.22)$$

The total variance σ^2_Y can easily be estimated from the observed data, as well as the constant μ. The first

© OECD 2005

purpose of variance component analysis is to obtain an estimate of the three variance components σ_{α}^{2}, $\sigma_{|}^{2}$, and σ_{ε}^{2}. If the data matrix is complete, good estimators are given by the traditional techniques of variance analysis, using the decomposition of the total sum of squares SS_{tot} as:

$$SS_{tot} = SS_{row} + SS_{col} + SS_{res} \qquad (10.23)$$

where *row* refers to the students and *col* refers to the items. Dividing each *SS* by their respective number of degrees of freedom yields the corresponding so-called mean squares, from which unbiased estimates of the three unknown variance components can be derived:

$$\hat{\sigma}_{\varepsilon}^{2} = MS_{res} \qquad (10.24)$$

$$\hat{\sigma}_{\alpha}^{2} = \frac{MS_{row} - MS_{res}}{I} \qquad (10.25)$$

and

$$\hat{\sigma}_{|}^{2} = \frac{MS_{col} - MS_{res}}{V} \qquad (10.26)$$

Usually the exact value of the three variance components will be of little use, but their relative contribution to the total variance is useful. Therefore the variance components will are normally expressed as a percentage of the total variance (the sum of the components).

The estimators given by (10.24) through (10.26) have the attractive property that they are unbiased, but they also have an unattractive property: the results of the formulae in (10.25) and (10.26) can be negative. In practice, this seldom occurs, and if it does, it is common to change the negative estimates to zero.

Analysis of three-way tables: The two-facet crossed design

For the two-facet (item and marker) crossed design, the model is a straightforward generalisation of the case of two-way tables. The observed data are now represented by Y_{vim}, the score student *v* receives for an answer on item *i* when marked by coder *m*. The observed data are arranged in a three-dimensional array (a box), where the student dimension will be denoted as *rows*, the item dimensions as *columns* and the marker dimensions as *layers*.

The model is a generalisation of model (10.20):

$$Y_{vim} = \mu + \alpha_{v} + |_{i} + \gamma_{m} + (\alpha|)_{vi} + (\alpha\gamma)_{vm} + (|\gamma)_{im} + \varepsilon_{vim} \qquad (10.27)$$

The observed variable Y_{vim} is the sum of a constant, three main effects, three first-order interactions and a residual. The residual in this case is the sum of the second-order interaction $(\alpha|\gamma)_{vim}$ and the measurement error ε_{vim}^{*}. Both effects are confounded because there is only one observation in each cell of the three-dimensional data array. The same restrictions as in the case of a two-way table apply: zero mean of the effects and mutual independence. Therefore the total variance decomposes into seven components:

$$\sigma_{Y}^{2} = \sigma_{\alpha}^{2} + \sigma_{|}^{2} + \sigma_{\gamma}^{2} + \sigma_{\alpha|}^{2} + \sigma_{\alpha\gamma}^{2} + \sigma_{|\gamma}^{2} + \sigma_{\varepsilon}^{2} \qquad (10.28)$$

and each component can be estimated with techniques similar to those demonstrated in the case of a two-way table.

The risk of ending up with negative estimates is usually greater for a three-dimensional table than for a two-way table. The main effect γ_m reflects the relative leniency of marker *m*: marker *m* is relatively mild if the effect is positive; relatively strict if it is negative. It is not unrealistic to assume that markers differ systematically in mildness, meaning that the variance component σ_Y^2 will differ markedly from zero, and consequently that its estimator will have a very small probability of yielding a negative estimate. A positive interaction effect γ_{vm} means that coder *m* is especially mild for student *v* (more than on average towards the other students), reflecting in some way a positive or negative bias to some students. But if the coding procedure has been seriously implemented – students not being known to the coders – it is to be expected that these effects will be very small if they do exist at all. And this means that the corresponding variance component will be close to zero, making a negative estimate quite likely.

Analysis of three-way tables: The special nested two-facet design

The difference between crossed and nested designs could be illustrated by the following example: a number of young musicians have to play a number of fragments from different composers, and each performance has to be scored by a number of jury members. The fragments have the role of the items; the jury members act as markers. The whole contest could be arranged in two different ways:

- Each student plays each fragment only once in the presence of the whole jury (nested design); or
- Each student plays all fragments in turn for each jury member individually (crossed design).

In both cases the data collection will be arranged in a similar three-way table, and in both cases an analysis will be carried out in an identical way, but the interpretation of the variance components will be different. A truly crossed design will probably never occur in educational settings.

Two sources of variability in the measurement error could be separated in the nested model. The model is split in a two-steps model: the first step models what happens when the student answers an item (with a given performance as an output), and the second step models what happens when a rater rates such a performance. So the output of the first step is the input of the second step, and the output of the second step is the observed item score by marker m: Y_{vim}.

The first step of the model is:

$$K_{vi} = M + A_v + B_i + (AB)_{vi} + E^*_{vi} \qquad (10.29)$$

Where *M* is the general effect, A_v is a main effect due to the student, B_i is a main effect due to an item, $(AB)_{vi}$ is an interaction effect of student and item and E^*_{vi} is a measurement error. The main effect, the interaction and the measurement error are considered as independent random variables with a mean of zero and with variances σ^2_A, σ^2_B, σ^2_{AB} and σ^2_E respectively. K_{vi} is a quantitative variable which is unobserved, but will be treated as a support variable.

In the second step, K_{vi} is amended by the marker to produce the observed score Y_{vim}. Such amending may be influenced by a main effect of the marker, or an interaction effect between marker and student or between marker and item, or a second order effect (marker by student by item) and an unsystematic effect, a measurement error (at the marker level). All these effects could be split into a mean effect (across markers,

students and items), and deviation from the mean, and all mean effects can be collected into a grand mean μ. So the second step is:

$$Y_{vim} = K_{vi} + m + b_i + c_m + (ac)_{vm} + (bc)_{im} + (abc)_{vim} + e^*_{vim} \qquad (10.30)$$

Replacing K_{vi} in the right-hand side of (10.30) by the right-hand side of equation (10.29), and grouping all the terms with the same set of subscripts results in:

$$\begin{aligned} Y_{vim} = & [M + m] \\ & + A_v + [B_i + b_i] + c_m \\ & + [(AB)_{vi} + E^*_{vi}] + (ac)_{vm} + (bc)_{im} \\ & + [(abc)_{vim} + e^*_{vim}] \end{aligned} \qquad (10.31)$$

Where M and m are constants, and all ten subscripted variables are random variables. It is impossible to estimate variances for the random variables with the same set of subscripts as they are confounded. But the variances of their sums could be estimated.

There are three pairs of the confounded variables in (10.31). One is the systematic item effect B_i which influences the unobserved variable K_{vi} and b_i which is a systematic item effect which comes about during coding of the performance. The second pair is the second order interaction effect and the measurement error at the marker level and the third is a confounding of the student-item interaction and the measurement error at the student level.

If the terms in the right-hand side of (10.31) are counted, counting bracketed terms as one single term, the result is one constant (first line), three main effects (second line), three first order interaction (third line) and a residual on the last line, which is just the same decomposition as in the genuine crossed design. This means that observed data can be arranged in the nested design in a three-way table which takes the same form as in the crossed design, and this table can be analysed in just the same way. The interpretation of the variance components, however, is different, as can be deduced from Table 10.2.

Table 10.2 ■ **Correspondence between the variance components in crossed and nested designs**

Crossed design		Nested design	
constant	μ	$[M+m]$	**constant**
persons	α_v	A_v	**persons**
items	β_i	$[B_i+b_i]$	**items**
coders	γ_r	c_r	**coders**
persons × items	$(\alpha\beta)_{vi}$	$[(AB)_{vi} + E^*_{vi}]$	**student × items + error at person level**
persons × coders	$(\alpha\gamma)_{vm}$	$(ac)_{vm}$	**student × coders**
items × coders	$(\beta\gamma)_{im}$	$(bc)_{im}$	**items × coders**
second order interactions + error	$\varepsilon_{vim} = [(\alpha\beta\gamma)_{vim} + \varepsilon^*_{vim}]$	$e_{vim} = [(abc)_{vim} + e^*_{vim}]$	**second order interactions + error at coder level**

Correlations

If one wants to determine the reliability of a test, one of the standard procedures is to administer the test a second time (under identical circumstances), and compute the correlation between the two sets of test scores. This correlation is, by definition, the reliability of the test. But if all variance components are known, this correlation can be computed from them. This is illustrated here for the case of a two-way table. To make the derivations easy, the relative test scores $Y_{v.}$ are used, defined as:

$$Y_{v.} = \frac{1}{I}\sum_{i} Y_{vi} \qquad (10.32)$$

Using model (10.18) and substituting in (10.32) the following is obtained:

$$Y_{v.} = \mu + \alpha_{v} + \frac{1}{I}\sum_{i} \rceil_{i} + \frac{1}{I}\sum_{i} (\alpha\rceil)_{vi} + \frac{1}{I}\sum_{i} \varepsilon^{*}_{vi} \qquad (10.33)$$

If the test is administered a second time using the same students and the same items, the relative scores on the repeated test will of course have the same structure as the right-hand side of (10.33), and, moreover, all terms will be identical with the exception of the last one, because the test replication is assumed to be independent. To compute the covariance between the two sets of test scores, observe that the mean item effects in both cases are the same for all students, meaning that this average item effect does not contribute to the covariance or to the variance. And because the measurement error is independent in the two administrations, it does not contribute to the covariance. So the covariance between the two series of test scores is:

$$\mathrm{cov}\left(Y_{v.}, Y'_{v.}\right) = \sigma^{2}_{\alpha} + \frac{\sigma^{2}_{\alpha\rceil}}{I} \qquad (10.34)$$

The variance of the test scores is equal in the two administrations:

$$\mathrm{var}\left(Y_{v.}\right) = \mathrm{var}\left(Y'_{v.}\right) = \sigma^{2}_{\alpha} + \frac{\sigma^{2}_{\alpha\rceil} + \sigma^{2}_{\varepsilon^{*}}}{I} \qquad (10.35)$$

Dividing the right-hand sides of (10.34) and (10.35) gives the correlation:

$$\rho_{1}\left(Y_{v.}, Y'_{v.}\right) = \frac{\sigma^{2}_{\alpha} + \frac{\sigma^{2}_{\alpha\rceil}}{I}}{\sigma^{2}_{\alpha} + \frac{\sigma^{2}_{\alpha\rceil} + \sigma^{2}_{\varepsilon^{*}}}{I}} \qquad (10.36)$$

But this correlation cannot be computed from a two-way table, because the interaction component $\sigma^{2}_{\alpha\rceil}$ cannot be estimated. It is common to drop this term from the numerators in (10.36), giving as a result Cronbach's alpha coefficient (Sirotnik, 1970):

$$alpha = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{I}} \leq \rho_1\left(Y_{v\cdot}, Y'_{v\cdot}\right) \qquad (10.37)$$

where the equality holds if and only if the interaction component $\sigma_{\alpha|}^2$ is zero.

In generalisability theory, Cronbach's alpha is also called the generalisability coefficient for relative decisions, meaning that it does not matter if one uses easy or difficult items to rank people, because in either case the relative standing of two persons with respect to each other will remain the same (within the bounds of measurement error).

Another interesting application arises when one wants to make a statement of a person's position with respect to the universe of items. An example might be an estimate of the proportion of words mastered in some (big) set of words. One could make such an estimate by administering a random sample of words to the test taker and considering the proportion of mastered words (in the sample) as an estimate of the proportion in the universe. To interpret the generalisability coefficient as a correlation coefficient, one assumes that for every test taker a new random sample is drawn, and the proportion correct is the average test score. In the retesting paradigm, this procedure is repeated a second time (so in general every test taker takes two different tests), and the generalisability coefficient for absolute decisions is the correlation between the two series of test scores. It is easily understood that the person-item interaction will not contribute to the covariance, because everybody takes two (independently drawn) tests, and moreover, the differences in difficulty between the tests now will contribute to the variance. The coefficient is given by:

$$\rho_2\left(Y_{v\cdot}, Y'_{v\cdot}\right) = \frac{\sigma_\alpha^2 + \frac{\sigma_{\alpha|}^2}{I}}{\sigma_\alpha^2 + \frac{\sigma_{|}^2 + \sigma_{\alpha|}^2 + \sigma_{\varepsilon^*}^2}{I}} \qquad (10.38)$$

It is easy to see that this coefficient cannot be greater than alpha (with equality only if the variance of the item effects is zero, *i.e.* if all items are equally difficult). Combining this with (10.38), an interesting relationship is revealed:

$$\rho_2 \leq alpha \leq \rho_1 . \qquad (10.39)$$

In the PISA study, where there are different sources of variation – three main effects, three first-order interactions and a residual effect – an alternative form of the generalisability coefficient needs to be derived.

Generalisability coefficients will be derived for two cases: one test administration where the performances are rated twice, each time by an independent sample of markers; and two independent administrations of the same test with each administration is rated by an independent set of *M* markers, randomly drawn from the universe of markers.

The relative test score is now defined as:

$$Y_{v\cdot\cdot} = \frac{1}{I \times M} \sum_i \sum_m Y_{vim} \qquad (10.40)$$

Of the eight terms on the right-hand side of (10.31), some are to be treated as constant, some contribute to the score variance and some to the covariance and the variance, as displayed in Table 10.3 for the two cases.

Table 10.3 ■ **Contribution of model terms to (co)variance in nested design**

	1	2
Constant	M, m, B, b, c, (bc)	M, m, B, b, c, (bc)
Variance & covariance	A, (AB), E^*	A, (AB)
Variance	e^*, (ac), (abc)	E^*, e^*, (ac), (abc)

Using this table, and taking the definition of the relative score into account, it is not too difficult to derive the correlation between the two series of test scores for the nested design (*i.e.* the first case):

$$\rho_3\left(Y_{v..}, Y'_{v..}\right) = \frac{\sigma_A^2 + \frac{\sigma^2_{AB+E^*}}{I}}{\sigma_A^2 + \frac{\sigma^2_{AB+E^*}}{I} + \frac{\sigma^2_{ac}}{R} + \frac{\sigma^2_{abc+e^*}}{I \times R}} \quad . \qquad (10.41)$$

For the second case formula is given by:

$$\rho_4\left(Y_{v..}, Y'_{v..}\right) = \frac{\sigma_A^2 + \frac{\sigma^2_{AB}}{I}}{\sigma_A^2 + \frac{\sigma^2_{AB} + \sigma^2_{\varepsilon^*}}{I} + \frac{\sigma^2_{ac}}{R} + \frac{\sigma^2_{abc+e^*}}{I \times R}} \qquad (10.42)$$

and it is immediately seen that the interaction component needed in the numerator is not available. In PISA, the Rasch model has been used as the IRT model, and this model presupposes absence of interaction between students and items. It is quite reasonable to assume that the component "student by item interaction plus error at the student level" is to be attributed to measurement error at the student level. Or, in other words, that the student by item component is zero.

The expressions on the right-hand side of (10.41) and (10.42) can be used as generic formulae for computing the correlation for an arbitrary number of items, and an arbitrary number of markers.

Estimation with missing data and incomplete designs

As described in the *PISA 2000 Technical Report* (OECD, 2002), the incomplete design used in PISA (see Chapter 6) introduces complications with respect to the estimation of the variance components that are needed to compute the reliabilities given in (10.41) and (10.42). The only commercial package known that can handle such problems is BMDP (1992), but the number of cases it can handle is very limited, and it was found to be inadequate for the PISA 2000 and 2003 analyses. Because of this it was necessary to develop and use a piece of special purpose software that used the moment-based algorithm described in the *PISA 2000 Technical Report* (OECD, 2002).

The variance components for the mathematics domain are displayed in Table 10.4 for Australia. A number of comments are required with respect to this table.

Table 10.4 ■ **Variance components (%) for mathematics in Australia**

Booklet	Number of Items	Student	Item	Coder	Student-Item Interaction Error	Student-Coder Interaction	Item-Coder Interaction	Measurement Error
1	6	16.9	22.8	0.00	56.9	0.0	0.1	3.3
2	6	20.2	18.1	0.01	56.4	0.0	0.0	5.3
3	5	28.2	3.4	-0.01	61.7	-0.1	0.1	6.7
4	8	14.6	38.9	0.00	40.3	0.0	0.1	6.0
5	11	18.1	37.9	0.02	38.9	0.1	0.1	4.9
6	10	9.8	33.1	0.01	49.4	-0.1	0.0	7.8
All	30	17.1	30.4	0.01	46.7	0.0	0.1	5.7

- The table is exemplary for all analyses in the four domains for countries where the coding instructions were rigorously followed. A more complete set of results is given in Chapter 14.
- The most comforting finding is that all variance components where coders are involved are negligibly small, meaning that there are no systematic marker effects.
- Using the bottom row of Table 10.5, the generalisability coefficients (10.41) and (10.42) can be computed for different values of *I* and *M*. These estimates are displayed, using data from the Australia, in Table 30. The result is exemplary for all countries, although the indices for reading and science are generally slightly lower than for mathematics and problem solving. But the main conclusion is that the correlations are quite high, even with one marker, such that the decision to use a single marker for the open-ended items in the main study consequently seems justified. Extensive tables for all participating countries are given in Chapter 14.

Table 10.5 ■ **Generalisability coefficients for mathematics scale for Australia**

I = 8				I = 16				I = 24			
M = 1	**M = 2**	**M = 1**	**M = 2**	**M = 1**	**M = 2**	**M = 1**	**M = 2**	**M = 1**	**M = 2**	**M = 1**	**M = 2**
0.969	0.984	0.722	0.733	0.982	0.991	0.838	0.846	0.987	0.993	0.886	0.89

INTER-COUNTRY CODER RELIABILITY STUDY DESIGN

As part of the PISA quality control procedures, reliability studies were conducted both in PISA 2000 and PISA 2003 in order to investigate the possibility of systematic bias in the coding of cognitive open-ended items used in the assessment. The within-country multiple coding exercise explored the reliability of the coding undertaken by the national coders in each country. The objective of the inter-country coder reliability (ICR) study was to check whether there was consistency between countries in the coding of open-ended items. Of particular interest was variation between countries in the level of severity of the coders.

The material used for the ICR study in PISA 2003 was a subset of 180 booklets (60 each of booklets 5, 8 and 10), randomly drawn from the sample that had been included in the multiple-coding exercise within each country. Booklet 5 was chosen because it began with two mathematics clusters, booklet eight was chosen because it began with two science clusters and booklet ten was chosen because it began with two Reading clusters, and in each case the first two clusters in the booklets contained ten extended open-ended questions. The coding consistency could therefore be checked using approximately 600 student responses

for mathematics, 600 for science and 600 for reading (10 items × 60 students in each domain) for each country. No problem-solving material was included in the study, for two reasons: no trend indicators had to be produced for problem solving; and the within-country homogeneity analyses undertaken at the field trial had shown extremely high rates of consistency among national coders, suggesting that international verification was of lesser importance for this domain than for the three others.

All participating countries were requested to submit to the consortium these 180 student booklets, after obscuring any codes given by the national coders. In countries where large percentages of the sampled students were assessed in more than one national language (*e.g.* Switzerland, Belgium, Latvia), the selection of booklets to be submitted was made in such a way that each language was appropriately represented in the sub-sample. In countries where only very small groups of students were assessed in a minority language (such as Hungarian in Serbia and Slovakia), only booklets in the dominant language were selected. In Spain, where separate adjudication of the data collected in the Basque and Catalan regions was needed, the ICR study was conducted using four separate sub-samples of 180 booklets (one for the Basque region, one for the Catalan region, one for the Castilla y Léon region and one for the rest of the country). Macao-China and Liechtenstein had their booklets coded, respectively, by Hong Kong-China and Switzerland; therefore, no separate ICR exercise was conducted for these two countries.

Staff specially trained by the consortium for the study, and proficient in the various PISA languages, then coded the booklets. Their codes were compared to the four codes given by the national staff to the same students' answers. All cases of clear discrepancy (between the code given by the independent coder and that given by all or most of the national coders) were submitted for adjudication consortium staff involved in developing the test materials, along with the actual student's answer (translated for all countries using languages other than English).

Recruitment of international coders

The booklets from the seven English-speaking PISA countries were coded and adjudicated at ACER (with England and Scotland being considered as separate countries for this exercise). To cover the languages of instruction used in all other PISA countries, the consortium appointed 22 of the translators who had already served in the verification of the national translations of the PISA 2003 material, and were therefore very familiar with it. The selection criteria for international coders were:

- Command of the language(s) of the country (or countries) whose material had to be scored. As far as possible the persons were chosen from among the verifiers who mastered more than one PISA language, so that each could code for two or more countries;
- Command of English and/or French, so that each of them could do the coding on the basis of either source version of the coding guides (while checking, when needed, the national version); and,
- As far as possible, previous experience in teaching either the national language or foreign languages at the secondary school level. In all cases when a verifier had insufficient background in some of the subject matter domains, the coding work was split between two different people (one person for reading and one for mathematics and science).

ICR training session

All international coders attended a three-day ICR training session conducted by consortium staff in Louvain-La-Neuve (Belgium).

The session materials included:

- The English and French source versions of the 30 items selected for the ICR study, with their coding instructions as presented in the source coding guides;
- Copies of the workshop materials used in the NPMs training sessions for these items;
- A selected list of answers to coding queries received at ACER for these items;
- The sets of ICR booklets received from the target countries; spreadsheets containing student and item identifiers, where the international coders were instructed to enter their codes (and, when needed, their translation of the students answers); and
- Copies of the national version of the coding guides used in the target countries.

During the session, the verifiers worked through the material question by question. Coding instructions for each question were presented, then the workshop examples were coded. Problem responses were discussed with the consortium staff conducting the training session, before proceeding to the next question. The verifiers were then instructed on how to enter their codes in the ICR study software prepared at ACER to help compare their codes with those given by the national coders. Most of them had the opportunity to start coding part of the material for one of the countries they were in charge of, under the supervision of consortium staff.

After completing a blind coding of each set of booklets, the verifiers received a spreadsheet indicating the cases where their scores differed from those given by the national coders. They went back to the booklets and reviewed their coding by entering their final code in another column. Cases where the international coders changed their code needed to be justified. They also entered a translation into English of the student's answer in a column next to each of the flagged cases. These so-called flag files were then returned to the ICR co-ordinators for adjudication.

Flag files

For each country, the ICR study software produced a flag file with about 1 800 coded responses (that is, 60 students × 30 items) – numbers varied slightly, since a few countries couldn't retrieve one or two of the booklets requested, or submitted booklets where one page was missing. In addition, for countries with booklets in different languages, two different coders were needed, and separate flag files were returned for each language subset.

In the flag file, a RU flag code systematically indicated cases where the verifier's code differed significantly from the four codes given by the national coders, that is:

- All cases where all four national coders gave a code that was the same but that differed from the verifier's code (*e.g.* national codes were 1, 1, 1, and 1 and verifier's code was 0, or the reverse).
- All cases when three out of four national coders gave a code that was the same but that differed from the verifier's code (*e.g.* national codes were 1, 1, 0, and 1 and verifier's code was 0).
- All cases when the national coders disagreed in the codes given, but at least three of them gave codes that yielded either a higher or lower score than the code given by the verifier (*e.g.* national codes were 1, 2, 1, and 0 and verifier's code was 0).

Cases with minor discrepancies only (*i.e.* where the verifier agreed with three out of four national coders) were not flagged, nor were most of the cases with national codes too inconsistent to be compared to the

verifier's code (*e.g.* where national codes were 0, 0, 1 and 1, and the verifier's code was 1; or cases where the national codes were 0, 1, 2, and 2, and the verifier's code was 2). It was considered that these cases would be identified in the within-country reliability analysis but would be of less interest for the ICR study than cases with a more clear orientation towards leniency or harshness. However, in PISA 2003, it was decided to include in the software an additional flag rule, so that a random 15 per cent of these overly inconsistent cases received a 'RD flag' code.

In each file, blank columns were left for consortium staff to adjudicate flagged cases.

Adjudication

In the adjudication stage, consortium staff adjudicated all flagged cases to verify whether the codes given by the national coders or by the verifier were the correct ones (or whether both the national coders and the verifier had used wrong codes in these problematic cases).

For the English-speaking countries, the consortium's adjudicators could check the flagged cases by directly referring to the students' answers in the booklets. For the non-English countries, the verifiers had been instructed to translate into English all of the flagged student answers.

All files for the non-English countries were adjudicated twice. A first adjudication was entered in the file by the ICR co-ordinators. Then the final adjudication was undertaken at ACER by mathematics, science and reading test developers).

Adjudication of coding discrepancies identified in student booklets from English-speaking countries was based on the same process as that used for all other countries. The sample of booklets to be reviewed for the inter-country coder reliability study was identified in the same way. Staff specially trained by the consortium then coded those booklets (in this part of the study, those staff were chosen from the most reliable coders from the Australian and New Zealand coding teams). Discrepancies were identified using exactly the same rules as for other countries, and all instances of discrepancy were referred to the item development experts from the consortium for final adjudication. The data from the study were used to calculate coder-reliability estimates for the English-speaking countries.

Country reports

Each country received a report after the adjudication process. The report contained the following three sections:

- A summary table presenting the per cent of cases in each of the agreement, inconsistency, harshness and leniency categories (overall and by domain);
- A table presenting details for each of the 30 open ended items included in the ICR study (where items with 10 per cent or more of either too harsh or too lenient codes and items with 10 per cent or more of too inconsistent codes were highlighted); and
- More qualitative comments on those items where some trend towards systematic leniency or harshness in the national coding was observed (if any).

The overall outcomes of the ICR study are presented in Chapter 14.

11

Data Cleaning Procedures

This chapter presents the data cleaning steps implemented during the main survey of PISA 2003.

DATA CLEANING AT THE NATIONAL CENTRE

National project managers (NPMs) were required to submit their national data in *KeyQuest*, the generic data entry package developed by consortium staff and pre-configured to include the data entry forms, referred to later as instruments: the achievement test booklets 1 to 13 (together making up the cognitive data); the Une Heure (UH) booklet; multiple-coding sheets; School Questionnaire and Student Questionnaire instruments with and without the two international options (the Information Communication Technology (ICT) questionnaire and the Educational Career questionnaires); the study programme table (SPT); the list of schools; and the student tracking forms.

The data were verified at several points starting at the time of data entry. Validation rules (or range checks) were specified for each variable defined in *KeyQuest*, and a datum was only accepted if it satisfied that validation rule.[1] To prevent duplicate records, a set of variables assigned to an instrument were identified as primary keys. For the student test booklets, the stratum, school and student identifications were the primary keys.

Countries were requested to enter data into the student tracking form module before starting to enter data for tests or context questionnaires. This module, or instrument, contained the complete student identification, as it should have appeared on the booklet and the questionnaire that the student received at the testing session. When configured, *KeyQuest* instruments designed for student data were linked with the student tracking form so that warning messages appeared when data operators tried to enter invalid student identifiers or student identifiers that did not match a record in the form.

After the data entry process was completed, NPMs were required to implement some of the checking procedures implemented in *KeyQuest* before submitting data to the consortium, and to rectify any integrity errors. These included inconsistencies between: the list of schools and the School Questionnaire; the student tracking form and achievement test booklets; the student tracking form and the Student Questionnaire; the achievement test booklets; and the Student Questionnaire. Also, in the multiple-coding data they checked in order to detect instances of other than four duplicate coders per booklet.

NPMs were required to submit a questionnaire adaptation spreadsheet with their data, describing all changes they had made to variables of the questionnaire, including the additions or deletions of variables or response categories, and changes to the meaning of response categories. NPMs were also required to propose recoding rules where the national data did not match the international data.

DATA CLEANING AT THE INTERNATIONAL CENTRE

Data cleaning organisation

Data cleaning was a major component of the PISA 2003 quality control and assurance programme. It was of prime importance that the consortium detected all anomalies and inconsistencies in submitted data, and that no errors were introduced during the cleaning and analysis phases. To reach these high quality requirements, the consortium implemented dual independent processing.

Two data analysts developed the PISA 2003 data cleaning procedures independently from each other, one using the statistical software package SAS®, the other using SPSS®. At each step, the procedures were

considered complete only when their application to a fictitious database and to the first two PISA databases received from countries produced identical results and files.

Three data-cleaning teams, each consisting of two data analysts, shared the work of producing the cleaned national databases. A team leader was nominated within each team and was the only individual to communicate with the national centres.

DATA CLEANING PROCEDURES

Because of the potential impact of PISA results and the scrutiny to which the data were likely to be put, it was essential that no dubious records remained in the data files. During cleaning, as many anomalies and inconsistencies as possible were identified, and through a process of extensive discussion between each national centre and the consortium's data processing centre at ACER, an effort was made to correct and resolve all data issues. When no adequate solution was found, the contradictory data records were deleted.[2]

Unresolved inconsistencies in student and school identifications led to the deletion of records in the database. Unsolved systematic errors for a particular item were replaced by not applicable codes. For instance, if countries reported a mistranslation or misprint in the national version of a cognitive booklet, data for these variables were recoded as not applicable and were not used in the analyses. Finally, errors or inconsistencies for particular students and particular variables were replaced by not applicable codes.

National adaptations to the database

When data arrived at the consortium, the first step was to check the consistency of the database structure with the international database structure. An automated procedure was developed for this purpose. For each instrument the procedure identified deleted variables, added variables and modified variables – that is, variables for which the validation rules had been changed. This report was then compared with the information provided by the NPM in the questionnaire adaptation spreadsheet. Once all inconsistencies were resolved, the submitted data were recoded where necessary to fit the international structure. All additional or modified variables were set aside in a separate file so that countries could use these data for their own purposes, but they were not included in the international database.

Verifying the student tracking form and the list of schools

The student tracking form and the list of schools were central instruments because they contained the information used in computing weight, exclusion, and participation rates. The student tracking form contained all student identifiers, inclusion and participation codes, the booklet number assigned and some demographic data. The list of schools contained, among other variables, the PISA population size, the grade population size and the sample size.

These forms were submitted electronically. The data quality in these two forms and their consistency with the booklets and Student Questionnaire data were verified for:

- Consistency of inclusion codes and participation codes with the data in the test booklets and questionnaires;
- Consistency of the sampling information in the list of schools (*i.e.* target population size and the sample size) with the student tracking form;

- Accordance with required international procedures of within-school student sampling;
- Consistency of demographic information in the student tracking form (grade, date of birth and gender) with that in the booklets or questionnaires; and
- Consistency of students' study programme code as listed on the student tracking form with codes in the student questionnaire and the study programme table.

Verifying the reliability data

Cleaning procedures were implemented to check the following components of the multiple-coding design (see Chapter 6):

- Number of records in the reliability files;
- Number of records in the reliability files and the corresponding booklets;
- Marker identification consistency;
- Marker design;
- Selection of the booklets for multiple coding; and
- Extreme inconsistencies in the marks given by different markers (see Chapter 14).[3]

Verifying the context questionnaire data

The Student Questionnaire and School Questionnaire data underwent further checks after the recoding of the national adaptations. Invalid or suspicious student and school data were reported to and discussed with countries. Four types of consistency checks were done:

- *Non-valid sums:* For example, question SCQ04 in the School Questionnaire requested the school principal to provide information as a percentage. The sum of the values had to be 100.
- *Implausible values:* Consistency checks across variables within instruments combined with the information of two or more questions to detect suspicious data. These checks included:

  Computing ratios like numbers of students (SCQ02) and numbers of teachers (SCQ18), numbers of computers (SCQ09) and numbers of students (SCQ02). Outlying ratios were identified and countries were requested to check the validity of the numbers.

  Comparing the mother's completed education level in terms of the International Standard Classification of Education (ISCED, OECD, 1999b) categories for ISCED 3A (STQ11) and ISCED 5A or higher (STQ12). If the mother did not complete a certain ISCED level, she could not have completed 5A; and

  Comparing the father's completed education level similarly.
- *Outliers*: All approximately normally distributed numeric answers from the School Questionnaire were standardised, and outlier values (±3 standard deviations) were returned to national centres for checking.[4]
- *Missing data confusion*: Possible confusions between 8, 9, 98, 99, 998, 999, 90, 900 when they are valid codes and missing or invalid values were encountered during the data cleaning. Therefore, for all numerical variables, values close to the missing codes were returned to countries for verification.

PREPARING FILES FOR ANALYSIS

For each PISA participating country, several files were prepared for use in the analysis:

- A raw cognitive file with all student responses to all items for the four domains before coding;
- A processed cognitive file with all student responses to all items for the four domains after coding;
- A student file with all data from the Student Questionnaire and the two international options;
- A school file with data from the School Questionnaire (one set of responses per school);
- Weighting files – two files with information from the student tracking form and the list of schools file necessary to compute the weights; and
- Reliability files – nine files with recoded student answers and nine files with scores, were prepared to facilitate the reliability analyses.

Processed cognitive file

For a number of items in the PISA test booklets, a student's score on the item was determined by combining two or more of their responses. Most recoding required combining two answers into one or summarising the information from the complex multiple-choice items.

In the PISA test material, some of the open-ended mathematics and science items were coded with two digits, while all other items were coded with a single-digit mark. *ConQuest*, the software used to scale the cognitive data, requires items of the same length. To minimise the size of the data file, the double-digit items were recoded into one-digit variables using the first digit. All data produced through scoring, combining or recoding have a *T* added to their variable label.

For items omitted by students, embedded missing and non-reached missing items were differentiated. All consecutive missing values clustered at the end of each booklet were replaced by a non-reached code (*r*), except for the first value of the missing series. Embedded and non-reached missing items were then treated differently in the scaling.

Non-reached items for students who were reported to have left the session earlier than expected were considered not applicable in all analyses.

The Student Questionnaire

The student questionnaire file includes the data collected from the student questionnaire and the international options. If a country did not participate in the international options, not applicable codes were used for the omitted variables.

Father's occupation, mother's occupation and the student's expected occupation at age 30, which were each originally coded using ISCO, were transformed into the International Socio-Economic Index of Occupational Status (ISEI) (Ganzeboom *et al.,* 1992).

Question EC07 regarding school marks was provided in three formats: nominal, ordinal and numerical. The nominal option was used if the country provided data collected with question EC07b – that is, above or at the pass mark and below the pass mark. Data collected through question EC07a were coded into EC07b according to the pass marks provided by national centres. Some countries submitted data for EC07a

with ranges of 1-5, or 1-7 etc., while others reported student marks on a scale with maximum scores of 20 or 100. These data were recoded in categorical format if fewer than eight categories were provided (1-7) or in percentages if more than seven categories were provided.

Calculator use and effort variables from the cognitive booklets were added to the student questionnaire file. The school level variable SC07Q01 (instructional school weeks per school year for each study programme) was linked to the Student Questionnaire by their programme code.

The School Questionnaire

No modifications other than the correction of data errors, the addition of the country three-digit codes and the computation of school indices were made to the School Questionnaire file. All participating schools, *i.e.* any school for which at least one PISA-eligible student was assessed, have a record in the international database, regardless of whether or not they returned the School Questionnaire.

The weighting files

The weighting files contained the information from the student tracking form and from the list of schools. In addition, the following variables were computed and included in the weighting files.

- For each student, a participation indicator was computed. If the student participated in the cognitive session of the original or follow-up sessions, then the student is considered a participant. If the student only attended the student questionnaire session, then the student was not considered a participant.
- For each student, a scorable variable was computed. All eligible students who attended a test session are considered as scorable. Further, if a student only attended the Student Questionnaire session and provided data for the father's or mother's occupation questions, then the student was also considered scorable. Therefore, an indicator was also computed to determine whether the student answered the father's or mother's occupation questions.

A few countries submitted data with a grade national option. Therefore, two eligibility variables – PISA-eligible and grade-eligible – were also included in the student tracking form. These new variables were also used to select the records that were included in the international database, and therefore in the cognitive file and Student Questionnaire file. To be included in the international database, the student had to be both PISA-eligible and scorable. PISA students reported in the student tracking form as not eligible, no longer at school, excluded for physical, mental, or linguistic reasons, or absent, as well as students who had refused to participate in the assessment, were not included in the international database.

All non-PISA students, *i.e.* students assessed in a few countries for a national or international grade sample option, were excluded from the international database. When countries submitted such data to the consortium it was processed and the results and clean data were returned separately to the national centres.

The reliability files

One file was created for each domain and test booklet. The data from the reliability booklets were merged with those in the test booklets so that each student selected for the multiple-coding process appears four times in these files.

Notes

1 National centres could modify the configuration of the variables, giving a range of values that was sometimes reduced or extended from the one originally specified by the Consortium.

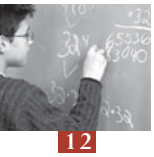
2 Record deletion was strenuously avoided as it decreased the participation rate.

3 For example, some markers reported a missing value while others reported non-zero scores.

4 The questions checked in this manner were school size (SCQ02), instructional time (SCQ07), number of computers (SCQ09), number of teachers (SCQ18 and SCQ19), number of hours per week spent on homework (STQ29), class size (STQ36) and school marks (ECQ07).

12

Sampling Outcomes



This chapter reports on PISA sampling outcomes. Details of the sample design are given in Chapter 4.

Table 12.1 shows the various quality indicators for population coverage, and the various pieces of information used to derive them. The following notes explain the meaning of each coverage index and how the data in each column of the table were used.

Indices 1, 2 and 3 are intended to measure PISA population coverage. Indices 4 and 5 are intended to be diagnostic in cases where indices 1, 2 or 3 have unexpected values. Many references are made in this chapter to the various sampling forms on which the National Project Managers (NPMs) documented statistics and other information needed in undertaking the sampling. The forms themselves are included in Appendix 1.

Index 1: Coverage of the national desired population, calculated by P/(P+E) ×3[c]/3[a].

- The national desired population (NDP), defined by sampling form 3 response box [a] and denoted here as 3[a] (and in Table 12.1 as "target desired population"), is the population that includes all enrolled 15-year-olds in each country in grades 7 and above (with the possibility of small levels of exclusions), based on national statistics. However, the final NDP reflected on each country's school sampling frame might have had some school-level exclusions. The value that represents the population of enrolled 15-year-olds minus those in excluded schools is represented initially by response box [c] on sampling form 3. It is denoted here as 3[c] (and in Table 12.1 as "target minus school level exclusions"). New in PISA 2003 was the procedure that very small schools having only one or two eligible students could not be excluded from the school frame, but could be excluded in the field if they still had exactly only one or two eligible students at the time of data collection. Therefore, what is noted in index 1 as 3[c] is a number that excludes schools excluded from the sampling frame in addition to those schools excluded in the field. Thus, the term 3[c]/3[a] provides the proportion of the NDP covered in each country based principally on national statistics.
- The value (*P*+*E*) provides the weighted estimate from the student sample of all eligible 15-year-olds in each country, where *P* is the weighted estimate of eligible non-excluded 15-year-olds and *E* is the weighted estimate of eligible 15-year-olds that were excluded within schools. Therefore, the term *P*/(*P*+*E*) provides an estimate based on the student sample of the proportion of the eligible 15-year-old population represented by the non-excluded eligible 15-year-olds.
- Thus, the result of multiplying these two proportions together (3[c]/3[a] and *P*/(*P*+*E*)) indicates the overall proportion of the NDP covered by the non-excluded portion of the student sample.

Index 2: Coverage of the national enrolled population, calculated by *P*/(*P*+*E*)×3[c]/2[b].

- The national enrolled population (NEP), defined by sampling form 2 response box [b] and denoted here as 2[b] (and as "enrolled 15-year olds" in Table 12.1), is the population that includes all enrolled 15-year-olds in each country in grade 7 and above, based on national statistics. The final NDP, denoted here as 3[c] as described above for coverage index 1, reflects the 15-year-old population after school-level exclusions. This value represents the population of enrolled 15-year-olds less those in excluded schools.
- The value (*P*+*E*) provides the weighted estimate from the student sample of all eligible 15-year-olds in each country, where *P* is the weighted estimate of eligible non-excluded 15-year-olds and *E* is the weighted estimate of eligible 15-year-olds that were excluded within schools. Therefore, the term *P*/(*P*+*E*) provides an estimate based on the student sample of the proportion of the eligible 15-year-old population that is represented by the non-excluded eligible 15-year-olds.

- Multiplying these two proportions together (3[c]/2[b] and $P/(P+E)$) gives the overall proportion of the NEP that is covered by the non-excluded portion of the student sample.

Index 3: Coverage of the national 15-year-old population, calculated by $P/2$[a].

- The national 15-year-old population, defined by sampling form 2 response box [a] and denoted here as 2[a] (called "all 15-year-olds" in Table 12.1), is the entire population of 15-year-olds in each country (enrolled and not enrolled), based on national statistics. The value P is the weighted estimate of eligible non-excluded 15-year-olds from the student sample. Thus, $P/2$[a] indicates the proportion of the national 15-year-old population covered by the eligible, non-excluded portion of the student sample.

Index 4: Coverage of the estimated school population, calculated by $(P+E)/S$.

- The value $(P+E)$ provides the weighted estimate from the student sample of all eligible 15-year-olds in each country, where P is the weighted estimate of eligible non-excluded 15-year-olds and E is the weighted estimate of eligible 15-year-olds that were excluded within schools.
- The value S is an estimate of the 15-year-old school population in each country (called "enrolled students on frame" in Table 12.1). This is based on the actual or (more often) approximate number of 15-year-olds enrolled in each school in the sample, prior to contacting the school to conduct the assessment. The S value is calculated as the sum over all sampled schools of the product of each school's sampling weight and its number of 15-year-olds (*ENR*) as recorded on the school sampling frame. In the infrequent case where the *ENR* value was not available, the number of 15-year-olds from the student tracking form was used.
- Thus, $(P+E)/S$ is the proportion of the estimated school 15-year-old population that is represented by the weighted estimate from the student sample of all eligible 15-year-olds. Its purpose is to check whether the student sampling has been carried out correctly, and to assess whether the value of S is a reliable measure of the number of enrolled 15-year-olds. This is important for interpreting Index 5.

Index 5: Coverage of the school sampling frame population, calculated by $S/3$[c].

- The value $S/3$[c] is the ratio of the enrolled 15-year-old population, as estimated from data on the school sampling frame, to the size of the enrolled student population, as reported on sampling form 3 and adjusted by removing any additional excluded schools in the field. In some cases, this provides a check as to whether the data on the sampling frame give a reliable estimate of the number of 15-year-olds in each school. In other cases, however, it is evident that 3[c] has been derived using data from the sampling frame by the NPM, so that this ratio may be close to 1.0 even if enrolment data on the school sampling frame are poor. Under such circumstances, Index 4 will differ noticeably from 1.0, and the figure for 3[c] will also be inaccurate.

Tables 12.2, 12.3 and 12.4 present school and student-level response rates. Table 12.2 indicates the rates calculated by using only original schools and no replacement schools. Table 12.3 indicates the improved response rates when first and second replacement schools were accounted for in the rates. Table 12.4 indicates the student response rates among the full set of participating schools.

For calculating school response rates before replacement, the numerator consisted of all original sample schools with enrolled age-eligible students who participated (*i.e.* assessed a sample of eligible students, and obtained a student response rate of at least 50 per cent). The denominator consisted of all the schools in the numerator, plus those original sample schools with enrolled age-eligible students that either did not participate or failed to assess at least 50 per cent of eligible sample students. Schools that were included in the sampling frame, but were found to have no age-eligible students, or which were excluded in the

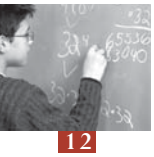


Table 12.1 ■ Sampling and coverage rates

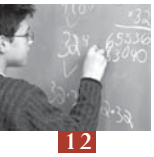
	All 15-year-olds	Enrolled 15-year-olds	Target desired population	School-level exclusions	Target minus school-level exclusions	School-level exclusions (%)	Enrolled students on frame	Participants		Excluded	
								Actual	Weighted	Actual	Weighted
Australia	268 164	250 635	248 035	1 621.00	246 414	0.65	275 208	12 551	235 591	228	3 612
Austria	94 515	89 049	89 049	320.59	88 728	0.36	87 795	4 597	85 931	60	1 099
Belgium	120 802	118 185	118 185	561.00	117 624	0.47	118 010	8 796	111 831	102	1 193
Brazil	3 618 332	2 359 854	2 348 405	0.00	2 348 405	0.00	2 340 538	4 452	1 952 253	5	2 142
Canada[9]	398 865	399 265	397 520	6 600.11	390 920	1.66	375 622	27 953	330 436	1 993	18 328
Czech Republic[11]	130 679	126 348	126 348	1 294.08	125 054	1.02	123 855	6 320	121 183	22	218
Denmark[14]	59 156	58 188	58 188	628.00	57 560	1.08	56 234	4 218	51 741	214	2 321
Finland[12]	61 107	61 107	61 107	1 324.00	59 783	2.17	59 766	5 796	57 883	79	725
France	809 053	808 276	774 711	18 056.00	756 655	2.33	757 355	4 300	734 579	51	8 158
Germany[15]	951 800	916 869	916 869	5 600.00	911 269	0.61	904 387	4 660	884 358	61	11 533
Greece[7, 13]	111 286	108 314	108 314	808.45	107 506	0.75	102 384	4 627	105 131	144	2 652
Hong Kong-China	75 000	72 631	72 631	601.00	72 030	0.83	72 312	4 478	72 484	8	103
Hungary	129 138	123 762	123 762	0.00	6 939	0.00	118 207	4 765	107 044	62	1 065
Iceland	4 168	4 112	4 112	3 687.54	120 074	2.98	4 086	3 350	3 928	79	79
Indonesia[5]	4 281 895	3 113 548	2 968 756	26.00	4 086	0.63	2 173 824	10 761	1 971 476	0	0
Ireland[17]	61 535	58 997	58 906	9 292.38	2 959 464	0.31	58 499	3 880	54 850	139	1 619
Italy[1]	561 304	574 611	574 611	864.43	58 042	1.47	563 039	11 639	481 521	188	6 794
Veneto - NE	37 843	36 388	36 388	2 868.48	571 743	0.50	35 056	1 538	30 854	22	416
Trento - NE	4 534	4 199	4 199	242.47	36 146	0.67	3 962	1 030	3 324	20	73
Toscana - Centro	27 111	29 208	29 208	76.85	4 122	1.83	28 272	1 509	25 722	21	346
Piemonte - NW	33 340	33 242	33 242	160.77	29 047	0.55	33 552	1 565	30 107	27	522
Lombardia - NW	76269	74 994	74 994	185.19	33 057	0.56	72 657	1 545	63 916	38	2 037
Bolzano - NE	4 908	4 087	4 087	252.11	74 742	0.34	3 967	1 264	3 464	25	67
Japan	1 365 471	1 328 498	1 328 498	9.12	4 078	0.22	1 314 227	4 707	1 240 054	0	0
Korea	606 722	606 370	606 370	2 729.00	603 641	0.45	614 825	5 444	533 504	24	2 283
Latvia	37 544	37 138	37 138	13 592.00	1 314 906	1.02	35 509	4 627	33 643	44	380
Liechtenstein	402	348	348	1 419.00	35 719	3.82	34 800	332	338	5	5
Luxembourg[16]	4 204	4 204	4 204	0.00	348	0.00	4 090	3 923	4 080	66	66
Macao-China	8 318	6 939	6 939	0.00	4204	0.00	6 992	1 250	6 546	4	13
Mexico[23]	2 192 452	1 273 163	1 273 163	46 482.97	1 226 680	3.65	1 204 851	29 983	1 071 650	34	7 264
Netherlands[3]	194 216	194 216	194 216	2 559.00	191 657	1.32	195 725	3 992	184 943	20	1 041
New Zealand	55 440	53 293	53 160	194.00	52 966	0.36	53 135	4 511	48 638	263	2 411
Norway	56 060	55 648	55 531	294.00	55 237	0.53	54 874	4 064	52 816	139	1 563
Poland	589 506	569 294	569 294	14 600.00	554 694	2.56	558 752	4 383	534 900	75	7 517
Portugal[8]	109 149	99 216	99 216	826.42	98 390	0.83	106 916	4 608	96 857	84	1 450
Russian Federation[10]	2 496 216	2 366 285	2 366 285	23 445.00	2 342 840	0.99	2 343 728	5 974	2 153 373	35	14 716
Serbia[6, 20]	98 729	92 617	92 617	4 931.17	87 686	5.32	90 178	4 405	68 596	15	241
Slovak Republic	84 242	81 945	81 890	1 042.00	80 848	1.27	80 626	7 346	77 067	109	1 341
Spain[1, 19]	454 064	418 005	418 005	1 639.00	416 366	0.39	412 829	10 791	344 372	591	25 619
Castilla-Leon	24 210	21 580	21 580	109.00	21 471	0.51	20 950	1 490	18 224	95	1 057
Catalonia	62 946	61 829	61 829	576.00	61 253	0.93	59 609	1 516	50 484	61	1 847
Basque Country	18 160	17 753	17 753	15.00	17 738	0.08	18 059	3 885	16 978	56	252
Sweden[2]	109 482	112 258	112 258	1 614.86	110 643	1.44	113 511	4 624	107 104	144	3 085
Switzerland	83 247	81 020	81 020	2 760.43	78 260	3.41	80 011	8 420	86 491	194	893
Thailand	927 070	778 267	778 267	7 597.00	770 670	0.98	770 109	5 236	637 076	5	563
Tunisia[4]	164 758	164 758	164 758	553.00	164 205	0.34	163 555	4 721	150 875	1	31
Turkey	1 351 492	725 030	725 030	5 328.10	719 702	0.73	719 702	4 855	481 279	0	0
United Kingdom	768 180	736 785	736 785	24 773.08	712 012	3.36	710 203	9 535	698 579	270	15 062
Scotland	65 913	63 950	63 950	917.00	63 033	1.43	62 814	2 723	58 559	39	715
United States	3 979 116	3 979 116	3 979 116	0.00	3 979 116	0.00	3 774 330	5 456	3 147 089	534	246 991
Uruguay	53 948	40 023	40 023	58.73	39 964	0.15	42 677	5 835	33 775	18	80

For notes, please see the end of the chapter.

Table 12.1 ■ Sampling and coverage rates (*continued*)

	Ineligible		Eligible		Within-school exclusions (%)	Overall exclusions (%)	Ineligible (%)	Coverage indices				
	Actual	Weighted	Actual	Weighted				1	2	3	4	5
Australia	562	7 886	15 733	239 203	1.51	2.15	3.30	0.98	0.97	0.88	0.87	1.12
Austria	146	2 159	6 306	87 030	1.26	1.62	2.48	0.98	0.98	0.91	0.99	0.99
Belgium	154	1 634	9 600	113 024	1.06	1.53	1.45	0.98	0.98	0.93	0.96	1.00
Brazil	334	137 164	4 876	1 954 395	0.11	0.11	7.02	1.00	0.99	0.54	0.84	1.00
Canada[9]	1 638	18 439	34 582	348 764	5.26	6.83	5.29	0.93	0.93	0.83	0.93	0.96
Czech Republic[11]	52	919	7 070	121 401	0.18	1.20	0.76	0.99	0.99	0.93	0.98	0.99
Denmark[14]	88	980	4 906	54 062	4.29	5.33	1.81	0.95	0.95	0.87	0.96	0.98
Finland[12]	32	303	6 314	58 608	1.24	3.38	0.52	0.97	0.97	0.95	0.98	1.00
France	66	10 490	5 026	742 737	1.10	3.40	1.41	0.97	0.93	0.91	0.98	1.00
Germany[15]	84	14 555	5 150	895 891	1.29	1.89	1.62	0.98	0.98	0.93	0.99	0.99
Greece[7, 13]	86	1 707	4 998	107 783	2.46	3.19	1.58	0.97	0.97	0.94	1.05	0.95
Hong Kong-China	91	1 370	4 974	72 587	0.14	0.97	1.89	0.99	0.99	0.97	1.00	1.00
Hungary	134	3 225	5 197	108 109	0.99	3.94	2.98	0.96	0.96	0.83	0.91	0.98
Iceland	104	104	4 003	4 007	1.97	2.59	2.60	0.97	0.97	0.94	0.98	1.00
Indonesia[5]	80	18 841	10 960	1 971 476	0.00	0.31	0.96	1.00	0.95	0.46	0.91	0.73
Ireland[17]	129	1 462	4 871	56 469	2.87	4.29	2.59	0.96	0.96	0.89	0.97	1.01
Italy[1]	355	18 559	12 595	488 315	1.39	1.88	3.80	0.98	0.98	0.86	0.87	0.98
Veneto - NE	27	526	1 662	31 270	1.33	1.99	1.68	0.98	0.98	0.82	0.89	0.97
Trento - NE	24	56	1 098	3 397	2.16	3.95	1.66	0.96	0.96	0.73	0.86	0.96
Toscana - Centro	41	609	1 638	26 068	1.33	1.87	2.33	0.98	0.98	0.95	0.92	0.97
Piemonte - NW	53	979	1 688	30 628	1.70	2.25	3.20	0.98	0.98	0.90	0.91	1.01
Lombardia - NW	44	1 929	1 658	65 953	3.09	3.41	2.92	0.97	0.97	0.84	0.91	0.97
Bolzano - NE	19	59	1 343	3 531	1.90	2.11	1 .68	0.98	0.98	0.71	0.89	0.97
Japan	19	4 699	4 951	1 240 054	0.00	1.02	0.38	0.99	0.99	0.91	0.94	1.00
Korea	67	6 493	5 533	535 787	0.43	0.87	1.21	0.99	0.99	0.88	0.87	1.02
Latvia	69	538	4 984	34 023	1.12	4.89	1.58	0.95	0.95	0.90	0.96	0.99
Liechtenstein	2	2	343	343	1.46	1.46	0.58	0.99	0.99	0.84	0.99	1.00
Luxembourg[16]	51	51	4 143	4 146	1.59	1.59	1.23	0.98	0.98	0.97	1.01	0.97
Macao-China	55	204	1 278	6 559	0.20	0.20	3.10	1.00	1.00	0.79	0.94	1.01
Mexico[23]	2 032	87 407	32 890	1 078 914	0.67	4.30	8.10	0.96	0.96	0.49	0.90	0.98
Netherlands[3]	46	1 942	4 547	185 984	0.56	1.87	1.04	0.98	0.98	0.95	0.95	1.02
New Zealand	337	3 056	5 582	51 049	4.72	5.07	5.99	0.95	0.95	0.88	0.96	1.00
Norway	38	429	4 789	54 380	2.87	3.39	0.79	0.97	0.96	0.94	0.99	0.99
Poland	15	1 440	5 476	542 417	1.39	3.91	0.27	0.96	0.96	0.91	0.97	1.01
Portugal[8]	305	5 581	5 321	98 307	1.47	2.30	5.68	0.98	0.98	0.89	0.92	1.09
Russian Federation[10]	69	22 994	6 288	2 168 089	0.68	1.66	1.06	0.98	0.98	0.86	0.93	1.00
Serbia[6, 20]	294	3 949	4 844	68 837	0.35	5.66	5.74	0.94	0.94	0.69	0.76	1.03
Slovak Republic	57	640	8 103	78 408	1.71	2.96	0.82	0.97	0.97	0.91	0.97	1.00
Spain[1, 19]	80	999	12 246	369 991	6.92	7.29	0.27	0.93	0.93	0.76	0.90	0.99
Castilla-Leon	5	58	1 695	19 281	5.48	5.96	0.30	0.94	0.94	0.75	0.92	0.98
Catalonia	7	234	1 695	52 331	3.53	4.43	0.45	0.96	0.96	0.80	0.88	0.97
Basque Country	60	275	4 128	17 231	1.46	1.55	1.59	0.98	0.98	0.93	0.95	1.02
Sweden[2]	35	764	5 114	110 189	2.80	4.20	0.69	0.96	0.96	0.98	0.97	1.03
Switzerland	144	1 731	9 086	87 384	1.02	4.39	1.98	0.96	0.96	1.04	1.09	1.02
Thailand	116	14 984	5 344	637 639	0.09	1.06	2.35	0.99	0.99	0.69	0.83	1.00
Tunisia[4]	312	9 596	4 903	150 906	0.02	0.36	6.36	1.00	1.00	0.92	0.92	1.00
Turkey	95	9 925	5 010	481 279	0.00	0.73	2.06	0.99	0.99	0.36	0.67	1.00
United Kingdom	422	26 177	12 303	713 641	2.11	5.40	3.67	0.95	0.95	0.91	1.00	1.00
Scotland	129	2 234	3 268	59 273	1.21	2.62	3.77	0.97	0.97	0.89	0.94	1.00
United States	261	124 279	7 337	3 394 080	7.28	7.28	3.66	0.93	0.93	0.79	0.90	0.95
Uruguay	622	2 635	6528	33855	0.24	0.38	7.78	1.00	1.00	0.63	0.79	1.07

For notes, please see the end of the chapter.



field, were omitted from the calculation of response rates. Replacement schools do not figure in these calculations.

In calculating weighted school response rates, each school received a weight equal to the product of its base weight (the reciprocal of its selection probability) and the number of age-eligible students enrolled, as indicated on the sampling frame.

With the use of probability proportional-to-size sampling, in countries with few certainty school selections and no over-sampling or under-sampling of any explicit strata, weighted and unweighted rates are very similar. Thus, the weighted school response rate before replacement is given by the formula:

$$\begin{matrix} weighted\ school\ response\ rate \\ before\ replacement \end{matrix} = \frac{\sum_{i \in Y} W_i E_i}{\sum_{i \in (Y \cup N)} W_i E_i} \qquad (12.1)$$

where Y denotes the set of responding original sample schools with age-eligible students, N denotes the set of eligible non-responding original sample schools, W_i denotes the base weight for school i, $W_i = 1/P_i$ where P_i denotes the school selection probability for school i, and E_i denotes the enrolment size of age-eligible students, as indicated on the sampling frame.

The weighted school response rate, after replacement, is given by the formula:

$$\begin{matrix} weighted\ school\ response\ rate \\ after\ replacement \end{matrix} = \frac{\sum_{i \in (Y \cup R)} W_i E_i}{\sum_{i \in (Y \cup R \cup N)} W_i E_i} \qquad (12.2)$$

where Y denotes the set of responding original sample schools, R denotes the set of responding replacement schools, for which the corresponding original sample school was eligible but was non-responding, N denotes the set of eligible refusing original sample schools which were not replaced, W_i denotes the base weight for school i, $W_i = 1/P_i$, where P_i denotes the school selection probability for school i, and for weighted rates, E_i denotes the enrolment size of age-eligible students, as indicated on the sampling frame.

For unweighted student response rates, the numerator is the number of students for whom assessment data were included in the results, less those in schools with between 25 and 50 per cent student participation. The denominator is the number of sampled students who were age-eligible, and not explicitly excluded as student exclusions, nor part of schools with student participation between 25 and 50 per cent. The exception is cases where countries applied different sampling rates across explicit strata.

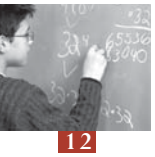
For weighted student response rates, the same number of students appears in the numerator and denominator as for unweighted rates, but each student was weighted by its student base weight. This is given as the product of the school base weight—for the school in which the student is enrolled—and the reciprocal of the student selection probability within the school.

In countries with no over-sampling of any explicit strata, weighted and unweighted student participation rates are very similar.

Overall response rates are calculated as the product of school and student response rates. Although overall weighted and unweighted rates can be calculated, there is little value in presenting overall unweighted

Table 12.2 ■ School response rates before replacements

	Weighted school participation rate before replacement (%)	Weighted number of responding schools *(also weighted by enrolment)*	Weighted number of schools sampled, responding and non-responding *(also weighted by enrolment)*	Unweighted school participation rate before replacement (%)	Number of Responding Schools *(unweighted)*	Number of responding and non-responding schools *(unweighted)*
Australia	86.31	237 525	275 208	84.79	301	355
Austria	99.29	87 169	87 795	98.97	192	194
Belgium	83.40	98 423	118 010	83.78	248	296
Brazil	93.20	2 181 287	2 340 538	93.01	213	229
Canada	79.95	300 328	375 622	89.50	1 040	1 162
Czech Republic	91.38	113 178	123 855	91.22	239	262
Denmark	84.60	47 573	56 234	83.33	175	210
Finland	97.39	58 209	59 766	97.97	193	197
France	88.65	671 417	757 355	88.52	162	183
Germany	98.06	886 841	904 387	97.69	211	216
Greece	80.60	82 526	102 384	81.01	145	179
Hong Kong-China	81.89	59 216	72 312	82.12	124	151
Hungary	97.32	115 041	118 207	94.66	248	262
Iceland	99.90	4 082	4 086	98.47	129	131
Indonesia	100.00	2 173 824	2 173 824	100.00	344	344
Ireland	90.24	52 791	58 499	90.26	139	154
Italy	97.54	549 168	563 039	98.03	398	406
Veneto – NE	97.97	34 344	35 056	98.08	51	52
Trento – NE	100.00	3 962	3 962	100.00	33	33
Toscana-Cntr	95.93	27 120	28 272	96.15	50	52
Piemonte-NW	96.12	32 249	33 552	96.49	55	57
Lombardia-NW	100.00	72 657	72 657	100.00	52	52
Bolzano - NE	100.00	3 967	3 967	100.00	43	43
Japan	87.12	1 144 942	1 314 227	87.33	131	150
Korea	95.89	589 540	614 825	95.97	143	149
Latvia	95.31	33 845	35 509	95.73	157	164
Liechtenstein	100.00	348	348	100.00	12	12
Luxembourg	99.93	4 087	4 090	90.63	29	32
Macao-China	100.00	6 992	6 992	100.00	39	39
Mexico	93.98	1 132 315	1 204 851	94.45	1 090	1 154
Netherlands	82.61	161 682	195 725	82.29	144	175
New Zealand	91.09	48 401	53 135	90.29	158	175
Norway	87.87	48 219	54 874	87.50	175	200
Poland	95.12	531 479	558 752	94.58	157	166
Portugal	99.31	106 174	106 916	99.35	152	153
Russian Federation	99.51	1 798 096	1 806 954	99.53	210	211
Serbia	100.00	90 178	90 178	100.00	149	149
Slovak Republic	78.92	63 629	80 626	78.52	223	284
Spain	98.39	406 170	412 829	98.43	377	383
Castilla-Leon	98.45	20 625	20 950	98.04	50	51
Catalonia	97.95	58 385	59 609	98.00	49	50
Basque Country	98.58	17 802.53	1 8059.02	98.58	139	141
Sweden	99.08	112 467	113 511	98.40	185	188
Switzerland	97.32	77 867	80 011	95.83	437	456
Thailand	91.46	704 344	770 109	91.06	163	179
Tunisia	100.00	163 555	163 555	100.00	149	149
Turkey	93.29	671 385	719 702	91.20	145	159
United Kingdom	64.32	456 818	710 203	68.96	311	451
Scotland	78.32	49 198	62 814	77.78	84	108
United States	64.94	2 451 083	3 774 330	65.18	249	382
Uruguay	93.20	39 773	42 677	95.10	233	245



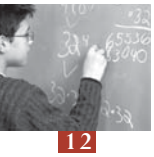
rates. The weighted rates indicate the proportion of the student population represented by the sample prior to making the school and student non-response adjustments.

Table 12.3 ■ **School response rates after replacement**

	Weighted school participation rate after replacement (%)	Weighted number of responding schools *(also weighted by enrolment)*	Weighted number of schools sampled, responding and non-responding *(also weighted by enrolment)*	Unweighted school participation rate after all replacement (%)	Number of responding schools *(unweighted)*	Number of responding and non-responding schools *(unweighted)*
Australia	90.43	248 876	275 208	88.45	314	355
Austria	99.29	87 169	87 795	98.97	192	194
Belgium	95.63	112 775	117 924	95.27	282	296
Brazil	99.51	2 328 972	2 340 538	99.56	228	229
Canada	84.38	316 977	375 638	91.74	1 066	1 162
Czech Republic	99.05	122 629	123 811	98.86	259	262
Denmark	98.32	55 271	56 213	97.62	205	210
Finland	100.00	59 766	59 766	100.00	197	197
France	89.24	675 840	757 355	89.07	163	183
Germany	98.82	893 879	904 559	98.61	213	216
Greece	95.77	104 859	109 490	95.53	171	179
Hong Kong-China	95.90	69 345	72 312	96.03	145	151
Hungary	99.37	117 269	118 012	96.18	252	262
Iceland	99.90	4 082	4 086	98.47	129	131
Indonesia	100.00	2 173 824	2 173 824	100.00	344	344
Ireland	92.84	54 310	58 499	92.86	143	154
Italy	100.00	563 039	563 039	100.00	406	406
Veneto – NE	100.00	35 056	35 056	100.00	52	52
Trento – NE	100.00	3 962	3 962	100.00	33	33
Toscana - Cntr	100.00	28 272	28 272	100.00	52	52
Piemonte – NW	100.00	33 552	33 552	100.00	57	57
Lombardia – NW	100.00	72 657	72 657	100.00	52	52
Bolzano – NE	100.00	3 967	3 967	100.00	43	43
Japan	95.91	1 260 428	1 314 227	96.00	144	150
Korea	100.00	614 825	614 825	100.00	149	149
Latvia	95.31	33 845	35 509	95.73	157	164
Liechtenstein	100.00	348	348	100.00	12	12
Luxembourg	99.93	4 087	4 090	90.63	29	32
Macao-China	100.00	6 992	6 992	100.00	39	39
Mexico	95.45	1 150 023	1 204 851	95.49	1 102	1 154
Netherlands	87.86	171 955	195 725	87.43	153	175
New Zealand	97.55	51 842	53 145	97.71	171	175
Norway	90.40	49 608	54 874	90.00	180	200
Poland	98.09	548 168	558 853	98.19	163	166
Portugal	99.31	106 174	106 916	99.35	152	153
Russian Federation	100.00	1 806 954	1 806 954	100.00	211	211
Serbia	100.00	90 178	90 178	100.00	149	149
Slovak Republic	99.08	80 394	81 141	98.94	281	284
Spain	100.00	412 777	412 777	100.00	383	383
Castilla-Leon	100.00	20 911	20 911	100.00	51	51
Catalonia	100.00	59 609	59 609	100.00	50	50
Basque Country	100.00	18 047	18 047	100.00	141	141
Sweden	99.08	112 467	113 511	98.40	185	188
Switzerland	98.53	78 838	80 014	97.39	444	456
Thailand	100.00	769 392	769 392	100.00	179	179
Tunisia	100.00	163 555	163 555	100.00	149	149
Turkey	100.00	719 405	719 405	100.00	159	159
United Kingdom	77.37	549 059	709 641	80.04	361	451
Scotland	88.89	55 737	62 794	88.89	96	108
United States	68.12	2 571 003	3 774 322	68.59	262	382
Uruguay	97.11	41 474	42 709	97.55	239	245

Table 12.4 ■ Student response rates after replacements

	Weighted student participation rate after replacements (%)	Number of students assessed *(weighted)*	Number of students sampled *(assessed + absent) (weighted)*	Unweighted student participation rate after replacements (%)	Number of students assessed *(unweighted)*	Number of students sampled *(assessed + absent) (unweighted)*
Australia	83.31	176 085.48	211 356.99	81.86	12 425	15 179
Austria	83.56	71 392.31	85 438.77	73.50	4 566	6 212
Belgium	92.47	98 935.93	106 994.65	92.61	8 796	9 498
Brazil	91.19	1 772 521.76	1 943 751.20	91.40	4 452	4 871
Canada	83.90	233 829.33	278 714.21	86.87	27 712	31 899
Czech Republic	89.03	106 644.57	119 791.10	89.77	6 316	7 036
Denmark	89.88	45 355.80	50 464.41	89.95	4 216	4 687
Finland	92.84	53 736.86	57 883.49	92.96	5 796	6 235
France	88.11	581 956.66	660 490.52	88.27	4 214	4 774
Germany	92.18	806 312.08	874 761.70	92.10	4 642	5 040
Greece	95.43	96 272.68	100 882.66	95.32	4 627	4 854
Hong Kong-China	90.20	62 755.77	69 575.73	90.17	4 478	4 966
Hungary	92.87	98 996.04	106 594.32	92.83	4 764	5 132
Iceland	85.37	3 350.00	3 924.00	85.37	3 350	3 924
Indonesia	98.09	1 933 838.77	1 971 476.30	98.18	10 761	10 960
Ireland	82.58	42 009.03	50 872.56	82.48	3 852	4 670
Italy	92.52	445 501.79	481 520.75	93.81	11 639	12 407
Veneto – NE	93.84	28 953.51	30 854.15	93.78	1 538	1 640
Trento – NE	95.97	3 189.69	3 323.75	95.55	1 030	1 078
Toscana - Cntr	93.04	23 930.56	25 722.08	93.32	1 509	1 617
Piemonte – NW	94.15	28 343.85	30 106.54	94.22	1 565	1 661
Lombardia – NW	95.48	61 024.16	63 915.67	95.37	1 545	1 620
Bolzano – NE	96.13	3 330.57	3 464.49	95.90	1 264	1 318
Japan	95.08	1 132 199.53	1 190 767.88	95.07	4 707	4 951
Korea	98.81	527 176.77	533 504.20	98.82	5 444	5 509
Latvia	93.88	300 42.86	32 001.41	93.66	4 627	4 940
Liechtenstein	98.22	332.00	338.00	98.22	332	338
Luxembourg	96.22	3 923.00	4 077.00	96.22	3 923	4 077
Macao-China	98.02	6 641.54	6 775.49	98.12	1 250	1 274
Mexico	92.26	938 901.78	1 017 666.73	92.12	29 734	32 276
Netherlands	88.25	144 211.88	163 417.98	88.46	3 979	4 498
New Zealand	85.71	40 595.43	47 362.84	85.67	4 483	5 233
Norway	87.86	41 922.64	47 714.86	87.92	4 039	4 594
Poland	81.95	429 920.50	524 583.62	81.91	4 338	5 296
Portugal	87.92	84 783.25	96 437.01	88.29	4 590	5 199
Russian Federation	95.71	2 061 050.06	2 153 373.33	95.54	5 974	6 253
Serbia	91.36	62 669.13	68 596.08	91.22	4 405	4 829
Slovak Republic	91.90	70 246.11	76 440.84	91.89	7 346	7 994
Spain	90.61	312 044.12	344 371.96	92.59	10 791	11 655
Castilla-Leon	93.28	16 999.74	18 223.90	93.13	1 490	1 600
Catalonia	92.95	46 922.34	50 483.51	92.78	1 516	1 634
Basque Country	95.38	16 194.83	16 978.49	95.41	3 885	4 072
Sweden	92.61	98 095.45	105 927.41	93.04	4 624	4 970
Switzerland	94.70	81 025.56	85 556.04	94.76	8 415	8 880
Thailand	97.81	623 092.96	637 075.68	98.07	5 236	5 339
Tunisia	96.27	145 250.92	150 874.89	96.31	4 721	4 902
Turkey	96.87	466 200.86	481 279.22	96.91	4 855	5 010
United Kingdom	77.92	419 810.06	538 737.19	81.62	9 265	11 352
Scotland	85.14	44 307.83	52 041.51	85.19	2 692	3 160
United States	82.73	1 772 279.24	2 142 287.58	82.16	5 342	6 502
Uruguay	90.83	29 755.57	32 759.39	90.27	5 797	6 422



DESIGN EFFECT AND EFFECTIVE SAMPLE SIZE

Surveys in education, and especially international surveys, rarely sample students by simply selecting a random sample of students (a simple random sample). Schools are first selected and, within each selected school, classes or students are randomly sampled. Sometimes, geographic areas are first selected before sampling schools and students. This sampling design is usually referred to as a cluster sample or a multi-stage sample.

Selected students attending the same school cannot be considered as independent observations, as they can be with a simple random sample because they are usually more similar than students attending distinct educational institutions. For instance, they are offered the same school resources, may have the same teachers and therefore are taught a common implemented curriculum, and so on. School differences are also larger if different educational programs are not available in all schools. One expects to observe greater differences between a vocational school and an academic school than between two comprehensive schools.

Furthermore, it is well known that within a country, within sub-national entities, and within a city, people tend to live in areas according to their financial resources. As children usually attend schools close to their house, it is likely that students attending the same school come from similar social and economic backgrounds.

A simple random sample of 4 000 students is thus likely to cover the diversity of the population better than a sample of 100 schools with 40 students observed within each school. It follows that the uncertainty associated with any population parameter estimate (*i.e.* standard error) will be larger for a clustered sample than for a simple random sample of the same size.

In the case of a simple random sample, the standard error on a mean estimate is equal to:

$$\sigma_{(\hat{\mu})} = \sqrt{\frac{\sigma^2}{n}} \qquad (12.3)$$

For an infinite population of schools and infinite populations of students within schools, the standard error of a mean estimate for a cluster sample is equal to:

$$\sigma_{(\hat{\mu})} = \sqrt{\frac{\sigma^2_{schools}}{n_{schools}} + \frac{\sigma^2_{within}}{n_{schools}\ n_{students}}} \qquad (12.4)$$

The standard error for a simple random sample is inversely proportional to the number of selected students. The standard error on the mean for a cluster sample is proportional to the variance that lies between clusters (*i.e.* schools) and within clusters and inversely proportional to the number of selected schools and the number of students selected per school.

It is usual to express the decomposition of the total variance into the between school variance and the within school variance by the coefficient of intraclass correlation, also denoted *rho*. Mathematically, this index is equal to

$$Rho = \frac{\sigma^2_{schools}}{\sigma^2_{schools} + \sigma^2_{within}} \qquad (12.5)$$

This index provides an indication of the percentage of variance that lies between schools.

© OECD 2005

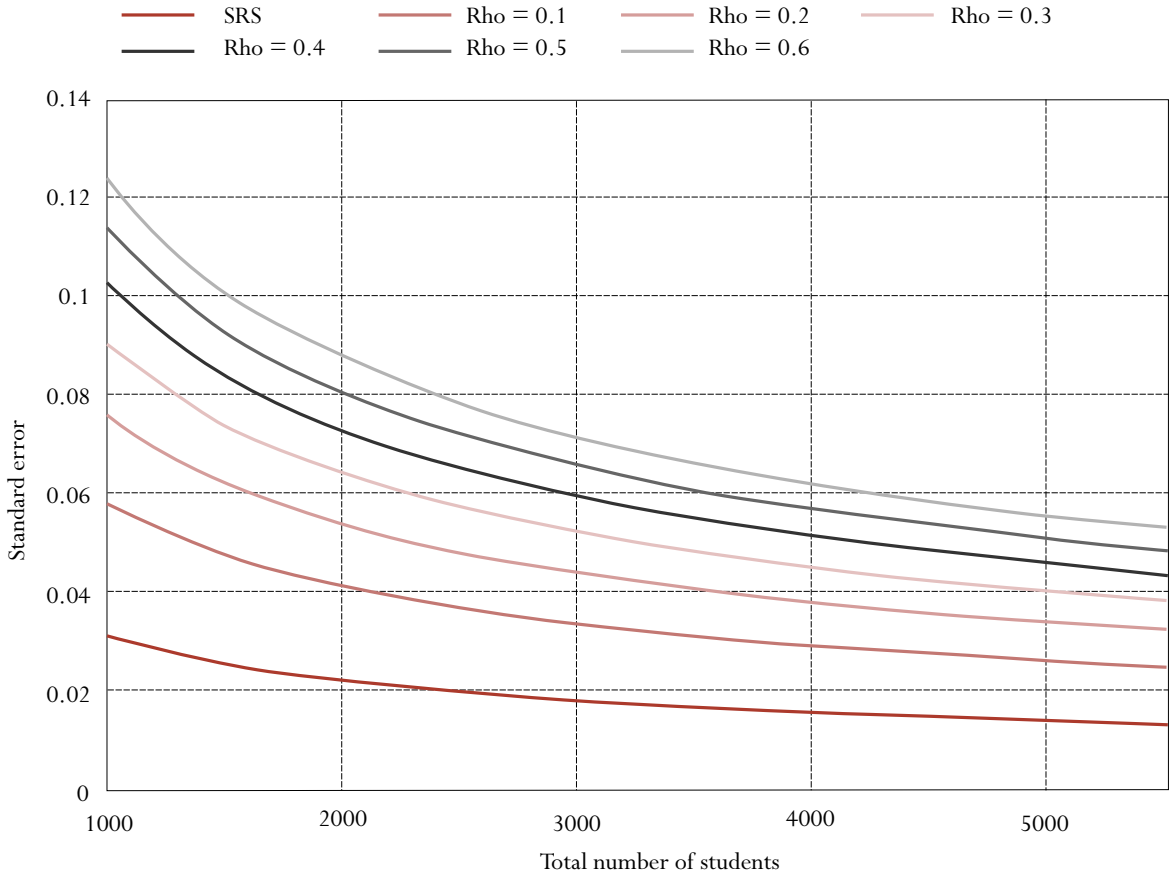
Figure 12.1 shows the standard errors on a mean for a simple random sample of 5 000 students and for cluster samples of 25 students per school for different intraclass correlation coefficients for any standardised variable. In the case of a sample of 25 students, it would mean that 200 schools would have participated.

Figure 12.1 shows that the standard error on the mean is quite a lot larger for a cluster sample than it is for a simple random sample and also that the standard error is proportional of the intraclass correlation.

To limit this reduction of precision in the population parameter estimate, multi-stage sample designs usually use complementary information to improve coverage of the population diversity. In PISA, and in previous international surveys, the following techniques were implemented to limit the increase in the standard error: *i*) explicit and or implicit stratification of the school sample frame, and *ii*) selection of schools with probabilities proportional to their size. Complementary information generally cannot compensate totally for the increase in the standard error due to the multi-stage design however.

Table 12.5 provides the standard errors on the PISA 2003 combined mathematical scale if the country sample was selected according to: *i)* a simple random sample; *ii)* a multistage procedure without using complementary information; and *iii)* the unbiased estimate using the Fay's replicates. It should be mentioned that the plausible value imputation variance was not included in these computations.

Figure 12.1 **Standard error on a mean estimate depending on the intraclass correlation**



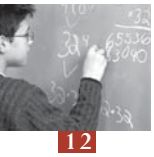


Table 12.5 ■ **Standard errors on the PISA 2003 mathematics scale**

	SRS	Cluster	Fay's BRR
Australia	0.85	2.63	2.13
Austria	1.37	5.39	3.23
Belgium	1.17	5.21	2.27
Brazil	1.49	4.46	4.78
Canada	0.52	1.34	1.78
Czech Republic	1.21	4.50	3.50
Denmark	1.41	2.75	2.66
Finland	1.10	1.79	1.78
France	1.40	4.88	2.46
Germany	1.50	5.42	3.31
Greece	1.38	4.83	3.88
Hong Kong-China	1.50	5.74	4.43
Hungary	1.35	5.13	2.77
Iceland	1.56	2.43	1.37
Indonesia	0.78	2.98	3.87
Ireland	1.37	3.18	2.40
Italy	0.89	3.68	2.97
Japan	1.47	6.23	3.99
Korea	1.25	5.06	3.18
Latvia	1.29	3.59	3.65
Liechtenstein	5.44	18.44	3.28
Luxembourg	1.47	9.92	0.96
Macao-China	2.46	6.82	2.83
Mexico	0.49	1.65	3.62
Netherlands	1.46	6.08	3.10
New Zealand	1.46	3.52	2.16
Norway	1.44	2.30	2.36
Poland	1.36	2.85	2.46
Portugal	1.29	4.32	3.40
Russian Federation	1.19	3.70	4.15
Serbia	1.28	4.31	3.69
Slovak Republic	1.09	3.80	3.32
Spain	0.85	2.32	2.35
Sweden	1.39	2.68	2.54
Switzerland	1.07	3.02	3.34
Thailand	1.13	3.98	2.94
Tunisia	1.19	4.45	2.52
Turkey	1.50	6.28	6.70
United Kingdom	0.94	2.61	2.38
United States	1.29	3.18	2.85
Uruguay	1.30	4.58	3.26

In several countries, the Fay's estimate of the standard error is substantially smaller than the estimate of the simple multistage sample. The difference provides an indication of the efficiency of the stratification process for reducing the sampling variance.

It is usual to express the effect of the sampling design on the standard errors by the design effect. It corresponds to the ratio of the variance of the estimate obtained from the (more complex) sample to the variance of the estimate that would be obtained from a simple random sample of the same number of units. The design effect has two primary uses – in sample size estimation and in appraising the efficiency of more complex plans (Cochran, 1977).

In PISA, as sampling variance has to be estimated by using the 80 Fay's replicate, a design effect can be computed for a statistic t using:

$$Deff\ (t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)} \qquad (12.6)$$

where $Var_{BRR}(t)$ is the sampling variance for the statistic t computed by the BRR replication method, and $Var_{SRS}(t)$ is the sampling variance for the same statistic t on the same data base but considering the sample as a simple random sample.

Based on the data of Table 12.5, the design effect in Australia for the mean estimate in mathematics is therefore equal to:

$$Deff\ (t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)} = \frac{(2.13)^2}{(0.85)^2} = 6.28 \qquad (12.7)$$

The sampling variance on the mathematics performance mean in Australia is about six times larger than it would have been with a simple random sample of equal size.

Another way to quantify the reduction of precision due to the complex sampling design is through the effective sample size, which expresses the simple random sample size that would give the same sampling variance as the one obtained from the actual complex sample design. The effective sample size for statistic t is equal to:

© OECD 2005

$$Effn(t) = \frac{n}{Deff(t)} = \frac{n \times Var_{SRS}(t)}{Var_{BRR}(t)} \qquad (12.8)$$

where n is equal to the actual number of units in the sample. The effective sample size in Australia for the mathematics performance mean is equal to:

$$Effn(t) = \frac{n}{Deff(t)} = \frac{n \times Var_{SRS}(t)}{Var_{BRR}(t)} = \frac{12551}{6.28} = 1999 \qquad (12.9)$$

In other words, a simple random sample of 1999 students in Australia would have been as precise as the actual PISA 2003 sample for the estimation of the mathematics performance.

Variability of the design effect

Neither the design effect nor the effective sample size are a definitive characteristic of a sample. Both depend on the requested statistic and on the variable on which some population parameters are estimated.

As stated previously, the sampling variance for a cluster sample is proportional to the intraclass correlation. In some countries, student performance varies between schools. Students in academic schools usually tend to perform well, while on average, student performance in vocational schools is lower. Let us now suppose that the height of the students was also measured. There are no reasons why students in academic schools should be taller than students in vocational schools, at least if there is no interaction between tracks and gender. For this particular variable, the expected value of the school variance should be equal to zero and therefore, the design effect should tend to one. As the segregation effect differs according to the variable, the design effect will also differ according to the variable.

The second factor that influences the size of the design effect is the requested statistics. It tends to be large for means, proportions, and sums but substantially smaller for bivariate or multivariate statistics such as correlations, regression coefficients and so on.

Design effects in PISA for performance variables

The notion of design effect as given earlier is extended and produces five design effects to describe the influence of the sampling and test designs on the standard errors for statistics.

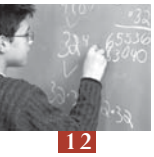
The total errors computed for the international PISA initial report that involves performance variables (plausible values or proficiency levels) consist of two components: sampling variance and measurement variance. The standard error in PISA is inflated because the students were not sampled according to a simple random sample and also because the measure of the student proficiency estimates includes some amount of random error.

For any statistic t, the population estimate and the sampling variance are computed for each plausible value and then combined as described in Chapter 9.

The five design effects, and their respective effective sample sizes, are defined as follows:

$$Deff_1(t) = \frac{Var_{SRS}(t) \quad MVar(t)}{Var_{SRS}(t)} \qquad (12.10)$$

where is the measurement variance for the statistic t. This design effect shows the inflation of the total



variance that would have occurred due to measurement error if in fact the sample were considered a simple random sample.

$$Deff_{2}(t) = \frac{Var_{BRR}(t) + MVar(t)}{Var_{SRS}(t) + MVar(t)} \qquad (12.11)$$

shows the inflation of the total variance due only to the use of the complex sampling design.

$$Deff_{3}(t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)} \qquad (12.12)$$

shows the inflation of the sampling variance due to the use of the complex design.

$$Deff_{4}(t) = \frac{Var_{BRR}(t) + MVar(t)}{Var_{BRR}(t)} \qquad (12.13)$$

shows the inflation of the total variance due to the measurement error.

$$Deff_{5}(t) = \frac{Var_{BRR}(t) + MVar(t)}{Var_{SRS}(t)} \qquad (12.14)$$

shows the inflation of the total variance due to the measurement error and due to the complex sampling design.

The product of the first and second design effects is equal to the product of the third and fourth design effects, and both products are equal to the fifth design effect.

Tables 12.6 to 12.8 provide the design effects and the effective sample sizes, respectively, for the country mean performance in mathematics, reading and science and the design effect for the percentage of students in the mathematic proficiency Level 3.

As previously mentioned, the design effects depend on the computed statistics. Except for Indonesia, Mexico and Turkey, the design effects are usually quite small.

Because the samples for the reading and science scales are drawn from the same schools as that for the combined mathematics scale, but with many fewer students, it follows that the mathematics sample is much more clustered than for the science and reading samples. Therefore it is not surprising to find that design effects are generally substantially higher for mathematics than for reading and science.

The measurement error for the minor domains is not substantially higher than the measurement error for the major domain because the proficiency estimates were generated with a multi-dimensional model using a large set of variables as conditioning variables. This complementary information has effectively reduced the measurement error for the minor domain proficiency estimates.

Table 12.6 ■ Design effects and effective sample sizes for the mean performance on the mathematical literacy scale

	Design effect 1	Design effect 2	Design effect 3	Design effect 4	Design effect 5	Effective sample size 1	Effective sample size 2	Effective sample size 3	Effective sample size 4	Effective sample size 5
Australia	1.11	5.75	6.26	1.02	6.36	11 335	2 184	2 006	12 339	1 973
Austria	1.14	4.97	5.52	1.02	5.66	4 040	924	833	4 485	812
Belgium	1.06	3.59	3.75	1.02	3.81	8 291	2 451	2 348	8 655	2 311
Brazil	1.22	8.54	10.23	1.02	10.45	3 639	521	435	4 357	426
Canada	1.51	8.08	11.67	1.04	12.17	18 559	3 458	2 396	26 791	2 296
Czech Republic	1.21	7.13	8.42	1.02	8.63	5 221	886	751	6 166	732
Denmark	1.24	3.07	3.57	1.07	3.81	3 402	1 373	1 182	3 952	1 108
Finland	1.25	2.30	2.63	1.10	2.88	4 626	2 519	2 204	5 288	2 011
France	1.12	2.87	3.09	1.04	3.21	3 851	1 498	1 392	4 143	1 342
Germany	1.01	4.81	4.86	1.00	4.87	4 603	968	959	4 648	957
Greece	1.10	7.25	7.89	1.01	8.00	4 192	639	586	4 567	579
Hong Kong-China	1.42	6.48	8.76	1.05	9.18	3 162	691	511	4 275	488
Hungary	1.20	3.66	4.19	1.05	4.39	3 978	1 301	1 137	4 550	1 086
Iceland	1.06	0.79	0.77	1.08	0.83	3 164	4 267	4 337	3 113	4 030
Indonesia	1.46	17.38	24.90	1.02	25.36	7 375	619	432	10 566	424
Ireland	1.11	2.87	3.09	1.04	3.20	3 483	1 351	1 258	3 742	1 213
Italy	1.78	6.77	11.24	1.07	12.02	6 556	1 719	1 035	10 888	968
Japan	1.09	6.87	7.42	1.01	7.51	4 308	685	635	4 649	627
Korea	1.22	5.48	6.47	1.03	6.69	4 457	994	842	5 264	814
Latvia	1.18	6.90	7.96	1.02	8.14	3 920	671	581	4 524	568
Liechtenstein	1.21	0.47	0.36	1.58	0.57	274	699	910	211	578
Luxembourg	1.01	0.43	0.43	1.03	0.44	3 872	9 055	9 215	3 805	8 937
Macao-China	1.05	1.31	1.33	1.04	1.38	1 189	955	943	1 204	908
Mexico	1.59	34.25	53.92	1.01	54.51	18 841	875	556	29 658	550
Netherlands	1.09	4.21	4.48	1.02	4.57	3 676	949	890	3 917	874
New Zealand	1.21	1.97	2.17	1.09	2.38	3 742	2 287	2 076	4 121	1 897
Norway	1.03	2.63	2.68	1.01	2.71	3 946	1 545	1 517	4 019	1 500
Poland	1.13	3.00	3.25	1.04	3.38	3 894	1 462	1 349	4 220	1 299
Portugal	1.02	6.84	6.94	1.00	6.96	4 534	673	664	4 597	662
Russian Federation	1.28	9.66	12.09	1.02	12.37	4 667	618	494	5 839	483
Serbia	1.29	6.73	8.38	1.03	8.66	3 424	654	526	4 259	508
Slovak Republic	1.14	8.32	9.32	1.01	9.45	6 466	883	788	7 240	777
Spain	1.36	5.87	7.64	1.05	8.00	7 918	1 838	1 413	10 302	1 348
Sweden	1.06	3.18	3.31	1.02	3.37	4 362	1 454	1 396	4 542	1 371
Switzerland	1.28	7.80	9.68	1.03	9.96	6 596	1 080	870	8 186	846
Thailand	1.25	5.59	6.75	1.04	7.01	4 177	937	775	5 047	747
Tunisia	1.05	4.30	4.47	1.01	4.52	4 497	1 097	1 057	4 669	1 045
Turkey	1.24	16.15	19.84	1.01	20.08	3 905	301	245	4 796	242
United Kingdom	1.26	5.25	6.34	1.04	6.60	7 588	1 816	1 504	9 164	1 446
United States	1.36	3.85	4.87	1.07	5.23	4 014	1 418	1 120	5 081	1 043
Uruguay	1.10	5.77	6.24	1.02	6.34	5 308	1 012	935	5 744	920

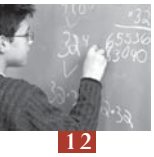


Table 12.7 ■ Design effects and effective sample sizes for the mean performance on the combined reading literacy scale

	Design effect 1	Design effect 2	Design effect 3	Design effect 4	Design effect 5	Effective sample size 1	Effective sample size 2	Effective sample size 3	Effective sample size 4	Effective sample size 5
Australia	1.22	4.92	5.77	1.04	5.99	10 328	2 548	2 175	12 100	2 097
Austria	1.10	5.58	6.02	1.02	6.11	4 195	824	764	4 525	752
Belgium	1.12	4.33	4.73	1.03	4.85	7 861	2 031	1 860	8 580	1 815
Brazil	1.37	5.49	7.17	1.05	7.54	3 244	810	621	4 232	591
Canada	1.49	7.29	10.39	1.05	10.89	18 723	3 833	2 690	26 687	2 568
Czech Republic	1.35	6.15	7.96	1.04	8.31	4 681	1 027	794	6 054	761
Denmark	1.39	3.09	3.90	1.10	4.30	3 032	1 366	1 080	3 834	982
Finland	1.16	2.06	2.22	1.07	2.38	5 009	2 820	2 609	5 413	2 437
France	1.16	2.83	3.12	1.05	3.28	3 707	1 522	1 379	4 090	1 312
Germany	1.05	4.29	4.44	1.01	4.49	4 454	1 087	1 050	4 612	1 039
Greece	1.52	4.70	6.60	1.08	7.12	3 054	985	701	4 292	650
Hong Kong-China	1.07	7.88	8.39	1.01	8.46	4 171	568	534	4 439	529
Hungary	1.12	3.08	3.32	1.03	3.43	4 271	1 548	1 436	4 605	1 388
Iceland	1.14	0.74	0.70	1.20	0.84	2 940	4 537	4 773	2 795	3 982
Indonesia	1.98	10.69	20.19	1.05	21.17	5 436	1 006	533	10 263	508
Ireland	1.13	3.16	3.44	1.04	3.57	3 434	1 228	1 127	3 739	1 086
Italy	1.90	5.59	9.73	1.09	10.63	6 123	2 081	1 196	10 653	1 095
Japan	1.31	4.97	6.20	1.05	6.51	3 595	947	759	4 483	723
Korea	1.24	6.14	7.39	1.03	7.63	4 379	887	737	5 271	713
Latvia	1.20	6.35	7.42	1.03	7.63	3 851	729	623	4 505	607
Liechtenstein	1.05	0.50	0.48	1.11	0.53	316	662	697	300	630
Luxembourg	1.36	0.64	0.51	1.70	0.87	2 890	6 121	7 654	2 311	4 509
Macao-China	1.29	1.01	1.01	1.28	1.30	970	1 236	1 233	973	960
Mexico	1.87	29.60	54.59	1.02	55.47	15 998	1 013	549	29 510	541
Netherlands	1.29	3.51	4.23	1.07	4.52	3 103	1 137	943	3 739	883
New Zealand	1.10	2.27	2.39	1.04	2.49	4 102	1 990	1 885	4 330	1 810
Norway	1.26	2.36	2.72	1.10	2.98	3 215	1 723	1 495	3 704	1 363
Poland	1.17	3.37	3.77	1.04	3.94	3 748	1 302	1 163	4 194	1 113
Portugal	1.11	6.75	7.36	1.01	7.46	4 166	683	626	4 543	617
Russian Federation	1.22	8.70	10.42	1.02	10.64	4 888	686	574	5 849	562
Serbia	1.11	7.59	8.30	1.01	8.41	3 977	580	530	4 349	524
Slovak Republic	1.03	8.10	8.33	1.00	8.37	7 111	907	882	7 317	878
Spain	1.83	4.38	7.19	1.12	8.02	5 898	2 463	1 502	9 674	1 346
Sweden	1.17	2.54	2.80	1.06	2.97	3 960	1 821	1 653	4 363	1 560
Switzerland	1.22	8.24	9.86	1.02	10.08	6 883	1 021	854	8 234	835
Thailand	1.70	3.97	6.06	1.12	6.76	3 073	1 320	865	4 691	775
Tunisia	1.48	2.74	3.58	1.14	4.06	3 181	1 726	1 320	4 158	1 163
Turkey	1.24	14.40	17.68	1.01	17.92	3 902	337	275	4 789	271
United Kingdom	1.47	4.46	6.09	1.08	6.56	6 489	2 137	1 567	8 852	1 455
United States	1.48	3.73	5.05	1.10	5.53	3 682	1 462	1 081	4 981	987
Uruguay	1.34	3.47	4.31	1.08	4.66	4 344	1 683	1 353	5 405	1 253

© OECD 2005

Table 12.8 ■ Design effects and effective sample sizes for the mean performance on the scientific literacy scale

	Design effect 1	Design effect 2	Design effect 3	Design effect 4	Design effect 5	Effective sample size 1	Effective sample size 2	Effective sample size 3	Effective sample size 4	Effective sample size 5
Australia	1.14	4.69	5.19	1.03	5.33	11 055	2 675	2 417	12 232	2 356
Austria	1.09	5.29	5.69	1.02	5.78	4 210	868	808	4 524	795
Belgium	1.47	3.18	4.20	1.11	4.67	5 987	2 767	2 093	7 912	1 883
Brazil	1.87	4.66	7.84	1.11	8.71	2 382	956	568	4 008	511
Canada	1.82	6.34	10.75	1.08	11.57	15 320	4 407	2 600	25 961	2 415
Czech Republic	1.58	4.52	6.55	1.09	7.12	4 006	1 400	965	5 808	887
Denmark	1.29	2.78	3.30	1.09	3.59	3 259	1 520	1 279	3 872	1 174
Finland	1.28	2.04	2.33	1.12	2.60	4 537	2 844	2 492	5 178	2 226
France	1.26	2.48	2.87	1.09	3.13	3 404	1 733	1 498	3 939	1 372
Germany	1.12	4.43	4.84	1.03	4.96	4 156	1 053	963	4 546	939
Greece	1.96	3.41	5.72	1.17	6.67	2 366	1 356	809	3 964	693
Hong Kong-China	1.19	7.74	8.99	1.02	9.18	3 777	578	498	4 387	488
Hungary	1.45	2.66	3.42	1.13	3.87	3 278	1 791	1 395	4 206	1 232
Iceland	1.05	0.75	0.74	1.07	0.79	3 179	4 469	4 551	3 122	4 240
Indonesia	1.70	14.11	23.26	1.03	23.95	6 340	762	463	10 448	449
Ireland	1.25	2.59	2.99	1.08	3.25	3 096	1 497	1 296	3 578	1 195
Italy	1.20	8.14	9.59	1.02	9.80	9 668	1 430	1 213	11 397	1 188
Japan	1.10	6.16	6.65	1.01	6.75	4 296	764	707	4 640	697
Korea	1.11	6.07	6.64	1.02	6.75	4 898	897	820	5 354	807
Latvia	1.15	7.08	7.99	1.02	8.14	4 026	654	579	4 542	569
Liechtenstein	1.16	0.50	0.42	1.39	0.58	285	665	795	238	571
Luxembourg	1.25	0.67	0.58	1.43	0.83	3 135	5 889	6 738	2 740	4 706
Macao-China	1.19	1.25	1.30	1.14	1.49	1 053	998	962	1 093	841
Mexico	5.90	8.22	43.61	1.11	48.51	5 078	3 649	688	26 952	618
Netherlands	1.29	3.15	3.78	1.08	4.07	3 093	1 267	1 057	3 707	981
New Zealand	1.16	2.00	2.15	1.07	2.31	3 891	2 261	2 094	4 201	1 950
Norway	1.14	2.73	2.97	1.05	3.11	3 570	1 487	1 367	3 883	1 306
Poland	1.04	3.30	3.39	1.01	3.43	4 222	1 328	1 293	4 334	1 279
Portugal	1.14	5.56	6.19	1.02	6.33	4 052	828	745	4 508	728
Russian Federation	1.15	8.92	10.14	1.02	10.29	5 178	670	589	5 885	580
Serbia	1.36	5.80	7.52	1.05	7.88	3 246	759	586	4 205	559
Slovak Republic	1.02	9.47	9.66	1.00	9.68	7 183	776	760	7 329	759
Spain	1.38	5.31	6.96	1.05	7.34	7 806	2 032	1 550	10 229	1 470
Sweden	1.43	2.11	2.59	1.17	3.01	3 240	2 191	1 789	3 968	1 535
Switzerland	1.20	8.26	9.69	1.02	9.89	7 033	1 019	869	8 252	851
Thailand	1.33	4.34	5.45	1.06	5.78	3 934	1 205	960	4 936	905
Tunisia	1.10	3.68	3.96	1.03	4.06	4 284	1 282	1 193	4 602	1 163
Turkey	1.26	14.56	18.04	1.01	18.29	3 864	333	269	4 787	265
United Kingdom	1.20	4.81	5.56	1.04	5.76	7 964	1 983	1 715	9 208	1 656
United States	1.32	3.80	4.69	1.07	5.01	4 139	1 437	1 164	5 109	1 090
Uruguay	1.04	3.95	4.07	1.01	4.11	5 608	1 478	1 435	5 778	1 421

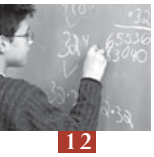


Table 12.9 ■ **Design effects and effective sample sizes for the percentage of students at Level 3 on the mathematical literacy scale**

	Design effect 1	Design effect 2	Design effect 3	Design effect 4	Design effect 5	Effective sample size 1	Effective sample size 2	Effective sample size 3	Effective sample size 4	Effective sample size 5
Australia	2.51	1.39	1.99	1.76	3.49	5 005	9 010	6 321	7 134	3 593
Austria	2.44	1.32	1.78	1.81	3.22	1 882	3 487	2 586	2 537	1 428
Belgium	2.00	1.39	1.78	1.56	2.78	4 406	6 319	4 935	5 643	3 166
Brazil	1.24	3.40	3.98	1.06	4.22	3 581	1 311	1 119	4 195	1 055
Canada	4.18	1.55	3.29	1.97	6.47	6 686	18 074	8 509	14 202	4 323
Czech Republic	1.24	2.48	2.84	1.08	3.08	5 107	2 543	2 227	5 832	2 055
Denmark	1.58	1.07	1.10	1.52	1.68	2 674	3 956	3 818	2 770	2 507
Finland	1.15	1.07	1.08	1.14	1.23	5 053	5 398	5 344	5 104	4 706
France	1.25	1.76	1.95	1.13	2.21	3 431	2 442	2 201	3 806	1 948
Germany	1.49	1.21	1.32	1.37	1.81	3 119	3 841	3 534	3 390	2 571
Greece	1.73	1.68	2.18	1.34	2.91	2 672	2 749	2 120	3 465	1 588
Hong Kong-China	3.44	1.27	1.92	2.28	4.36	1 301	3 538	2 338	1 968	1 028
Hungary	1.55	1.43	1.67	1.33	2.22	3 082	3 324	2 853	3 591	2 150
Iceland	1.39	0.97	0.96	1.40	1.35	2 418	3 444	3 482	2 392	2 486
Indonesia	1.88	5.63	9.69	1.09	10.57	5 729	1 912	1 110	9 867	1 018
Ireland	1.02	1.28	1.28	1.01	1.30	3 810	3 042	3 030	3 825	2 987
Italy	1.26	3.67	4.36	1.06	4.62	9 231	3 174	2 667	10 982	2 517
Japan	1.65	1.72	2.19	1.30	2.84	2 854	2 732	2 147	3 631	1 656
Korea	1.67	1.70	2.17	1.31	2.84	3 260	3 199	2 507	4 161	1 916
Latvia	2.29	1.38	1.88	1.69	3.17	2 021	3 345	2 464	2 743	1 461
Liechtenstein	1.21	1.05	1.06	1.20	1.27	275	316	313	277	261
Luxembourg	1.50	0.85	0.77	1.65	1.27	2 617	4 640	5 106	2 378	3 095
Macao-China	1.41	1.41	1.58	1.26	1.99	888	886	792	994	629
Mexico	3.31	7.00	20.87	1.11	23.17	9 062	4 281	1 437	26 996	1 294
Netherlands	1.55	1.88	2.36	1.23	2.91	2 582	2 123	1 691	3 242	1 373
New Zealand	1.99	1.03	1.07	1.92	2.06	2 269	4 360	4 220	2 344	2 193
Norway	2.00	1.10	1.21	1.83	2.20	2 035	3 684	3 370	2 224	1 845
Poland	1.71	1.19	1.33	1.53	2.04	2 564	3 680	3 304	2 856	2 153
Portugal	1.48	1.83	2.22	1.22	2.70	3 117	2 522	2 073	3 792	1 706
Russian Federation	1.56	2.24	2.94	1.19	3.50	3 818	2 669	2 034	5 011	1 706
Serbia	1.74	2.05	2.83	1.26	3.58	2 526	2 147	1 555	3 489	1 231
Slovak Republic	2.91	1.57	2.66	1.72	4.57	2 523	4 677	2 760	4 275	1 606
Spain	4.26	1.36	2.52	2.29	5.78	2 535	7 946	4 276	4 711	1 867
Sweden	2.01	1.09	1.18	1.85	2.19	2 306	4 234	3 903	2 501	2 111
Switzerland	1.36	3.25	4.05	1.09	4.41	6 204	2 591	2 077	7 738	1 909
Thailand	1.49	2.15	2.70	1.18	3.19	3 518	2 441	1 936	4 435	1 640
Tunisia	1.38	2.37	2.89	1.13	3.27	3 431	1 988	1 633	4 178	1 445
Turkey	2.10	3.19	5.59	1.20	6.68	2 316	1 523	869	4 059	726
United Kingdom	2.77	1.41	2.15	1.82	3.92	3 440	6 739	4 435	5 227	2 431
United States	1.48	1.29	1.43	1.33	1.90	3 696	4 232	3 824	4 091	2 867
Uruguay	1.13	1.71	1.80	1.07	1.93	5 157	3 413	3 236	5 438	3 016

Notes

1 The Italy and Spain entries are more than the sum of the listed parts since not all parts were required to be broken out.

2 Sweden's enrolled population is larger than the number of 15 year olds because it is based on estimated data from a different source.

3 The Netherlands' frame count of ENR was 196 908 because of rounding decimal values of ENR and imputing values of 1 when ENR was zero or missing.

4 Tunisia noted late in the process that one French school (121726) needed to be excluded because of French (rather than Arabic) language – it had 33 eligible students. This is reflected in the 3[b] number.

5 Indonesia excluded four provinces and close to 5 per cent of its eligible population due to security reasons. There were 4 137 103 15-years old for 2[a], but the four provinces were already excluded. Therefore, the 144 792 noted as being excluded in these provinces was added to this number to get 4 281 895 15-year-olds. The number of enrolled 15-year-olds was noted as 2968756 so 144 792 was also added to this. Then, the 14 4792 was taken off to arrive at the 3[a] number.

6 Serbia excluded Kosovo and there were no estimates for the number of 15-year-olds so this does not appear as an exclusion.

7 Greece originally had excluded students in primary schools but since the population was later changed to 15-year-olds in grades 7 and above, the population figures have been adjusted so that these are not exclusions, but not part of the population to begin with.

8 Portugal's enrolled number of 15-year-olds is likely an underestimate because this number came from schools that responded to questions about the number of 15-year-olds. There were non-respondents.

9 Canada's Sf2[b] is greater than the Sf3[a] number due to different data sources.

10 The Russian Federation's PSU frame is from 1999 statistics and had a frame count of 1 772 900 students, which likely underestimates the PISA 2003 population of 15-year-olds. Also, the school-level frame count was 1 422 600, which also likely underestimates the population over selected PSUs given an SF3[c] of 1 847 166 for the sampled regions only.

11 The Czech Republic's exclusion code 4 was for students abroad or absent for long periods. These students additionally had a SEN code for reading disorders.

12 Finland's exclusion code 4 was defined as dyslexia (after the fact).

13 Greece's exclusion code 4 was defined as dyslexia.

14 Denmark's exclusion code 4 was for dyslexia/acalculia.

15 Germany had six students excluded after the fact with code=4 after they were given the UH booklet in a school where not all students were given the UH booklet.

16 Luxembourg's exclusion code 4 was for students being "primo-arrivants". This code applies to students who have only very recently come to Luxembourg, normally as asylum-seekers.

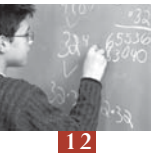
17 Ireland's exclusion code 4 was for dyslexia.

18 Poland's exclusion code 4 was for dyslexia.

19 Spain's sampling form numbers were updated from census figures for 2003.

20 Serbia originally had 724 for school-level exclusions. After weighting, it was realised that primary schools, although thought to be on the frame, were not. Thus, 3065 has been added to school-level exclusions.

21 To arrive at the adjusted column for SF2[b] with 15-year olds in grades 5 and 6 removed, one of 4 sources of country data were used for each country. For Australia, Brazil, Macao-China, Mexico, Thailand, Tunisia and Uruguay, sampling was done after the population definition so sampling forms numbers did not include counts for students in grades 5 and 6. Poland had these students as part of their school level exclusions-- they were removed from exclusions and used to arrive at the adjusted figure for SF2[b]. For Denmark, Germany, Italy, Latvia, Russia, Slovakia and Turkey, estimates from the sample were removed from the column for within-school student level exclusions for this reason, and used to adjust the original



SF2[b]. All other countries supplied estimates for adjusting SF2[b], except for Iceland and Luxembourg which did not supply any information so 0 in these grades has been assumed.

22 Canada did not have any ineligible students in grades 5 and 6. However they had excluded home school students under exclusion category 4, when really these students are being classed in other countries as ineligible. Thus these have been moved to ineligible grade 5/6 for Canada. Sampling form numbers have also been adjusted to remove the 66 students originally excluded from home schools. This was similarly done for the US (17 students and 8536 weighted).

23 Mexico could not conduct an assessment in the province of Michoacan (stratum 16) because of a teacher strike so all students in these schools have been regarded as exclusions at the school-level (46472 based on SF8).

13

Scaling Outcomes

INTERNATIONAL CHARACTERISTICS OF THE ITEM POOL

When main study data were received from each participating country, they were first verified and cleaned using the procedures outlined in Chapter 11. Files containing the achievement data were prepared and national-level Rasch and traditional test analyses were undertaken. The results of these analyses were included in the reports that were returned to each participant.

Table 13.1 ■ **Number of sampled students by country and booklet**

	Booklet														
	1	2	3	4	5	6	7	8	9	10	11	12	13	UH	Total
Australia	992	977	961	970	954	975	964	961	954	969	946	950	978		12 551
Austria	356	352	350	348	355	355	348	361	341	350	353	350	349	29	4 597
Belgium	660	658	649	644	645	647	660	661	656	665	657	667	680	247	8 796
Brazil	351	346	347	335	346	335	340	360	318	348	328	345	353		4 452
Canada	2 184	2 143	2 117	2 154	2 143	2 142	2 137	2 157	2 140	2 162	2 127	2 163	2 184		27 953
Czech Republic	466	464	468	472	483	482	497	478	487	469	473	472	481	128	6 320
Denmark	321	323	329	335	311	343	327	317	322	325	301	330	334		4 218
Finland	430	427	430	443	435	451	461	457	463	457	456	436	450		5 796
France	321	339	342	326	326	331	332	330	330	330	323	334	336		4 300
Germany	348	343	350	348	356	357	348	348	355	356	349	342	352	108	4 660
Greece	371	361	375	347	343	344	338	352	359	351	358	362	366		4 627
Hong Kong-China	349	348	339	344	349	348	345	346	331	351	339	344	345		4 478
Hungary	347	344	329	332	344	339	330	332	341	324	327	338	344	394	4 765
Iceland	254	258	256	261	263	259	259	265	261	251	255	254	254		3 350
Indonesia	817	817	826	830	815	820	810	831	842	848	846	834	825		10 761
Ireland	283	302	295	292	292	303	311	302	299	306	302	302	291		3 880
Italy	887	885	889	902	881	869	889	909	919	898	913	901	897		11 639
Japan	355	371	362	361	362	356	364	358	369	362	361	371	355		4 707
Korea	419	416	412	417	416	409	423	426	425	413	430	417	421		5 444
Latvia	359	363	357	356	360	358	357	358	354	348	353	358	346		4 627
Liechtenstein	28	26	23	27	25	26	23	25	23	27	25	27	27		332
Luxembourg	317	318	311	308	306	304	303	299	296	289	295	289	288		3 923
Macao-China	93	91	98	99	99	98	99	97	100	96	94	94	92		1 250
Mexico	2 321	2 304	2 327	2 319	2 318	2 330	2 294	2 308	2 296	2 293	2 298	2 272	2 303		29 983
Netherlands	299	315	308	288	299	306	296	309	296	297	298	299	298	84	3 992
New Zealand	339	352	347	335	338	338	342	342	343	347	353	376	359		4 511
Norway	310	309	314	320	314	322	310	308	299	310	321	316	311		4 064
Poland	349	342	339	330	318	325	340	342	331	345	331	356	335		4 383
Portugal	355	365	362	353	366	347	350	355	346	339	361	360	349		4 608
Russian Federation	461	469	472	473	470	449	455	455	439	461	456	462	452		5 974
Serbia	339	359	341	347	338	328	341	325	347	320	346	332	342		4 405
Slovak Republic	563	558	567	565	564	559	547	567	563	563	560	551	561	58	7 346
Spain	838	827	835	817	828	827	798	837	847	836	846	824	831		10 791
Sweden	368	368	372	369	364	362	350	354	333	342	344	344	354		4 624
Switzerland	634	665	652	646	648	649	649	663	646	639	647	629	653		8 420
Thailand	393	404	412	408	396	390	408	414	410	410	403	389	399		5 236
Tunisia	363	366	363	361	369	356	364	361	361	360	367	365	365		4 721
Turkey	383	391	379	364	378	381	380	375	370	365	366	363	360		4 855
United Kingdom	756	741	742	712	747	738	729	743	730	725	714	716	742		9 535
United States	418	425	420	408	431	427	407	418	421	437	424	404	416		5 456
Uruguay	462	467	455	457	439	450	455	447	446	442	435	441	439		5 835
Total	**21 259**	**21 299**	**21 222**	**21 123**	**21 134**	**21 135**	**21 080**	**21 253**	**21 109**	**21 126**	**21 081**	**21 079**	**21 217**	**1 048**	**276 165**

© OECD 2005

After processing at the national level, a set of international-level analyses was undertaken. Some involved summarising national analyses, while others required an analysis of the international data set.

The final international cognitive data set (that is, the data set of coded achievement booklet responses) (available as *intcogn.txt*) consisted of 276 165 students from 42 participating countries. Table 13.1 shows the total number of sampled students, broken down by participating country and test booklet.

Test targeting

Each of the domains was separately scaled to examine the targeting of the tests. Figures 13.1, 13.2, 13.3 and 13.4 show the match between the international item difficulty distribution and the international

Figure 13.1 ■ **Item plot for mathematics items**

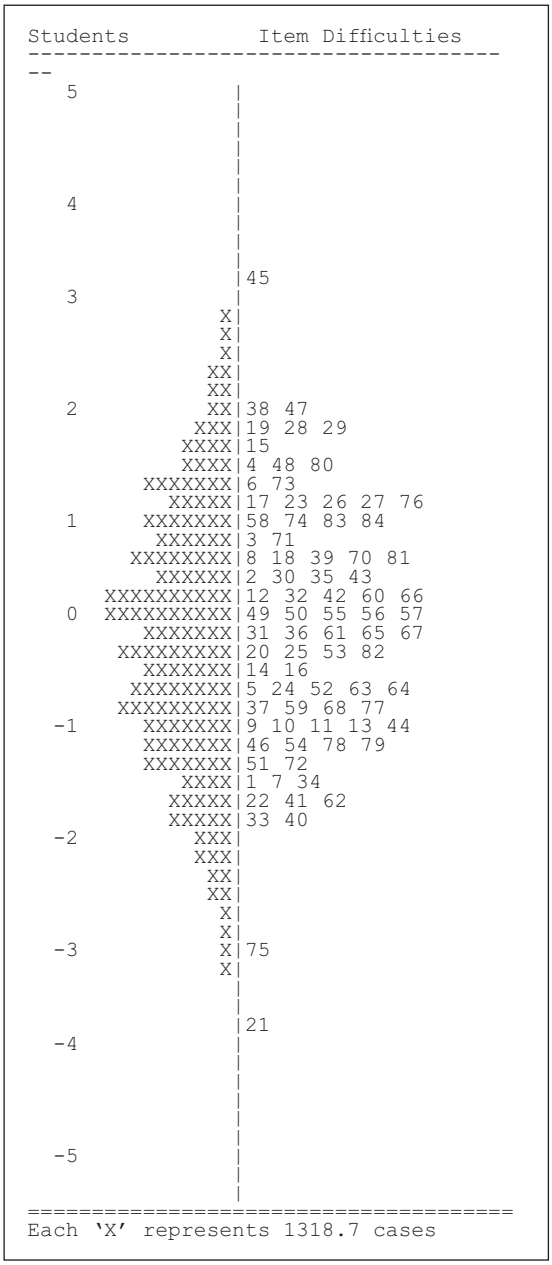
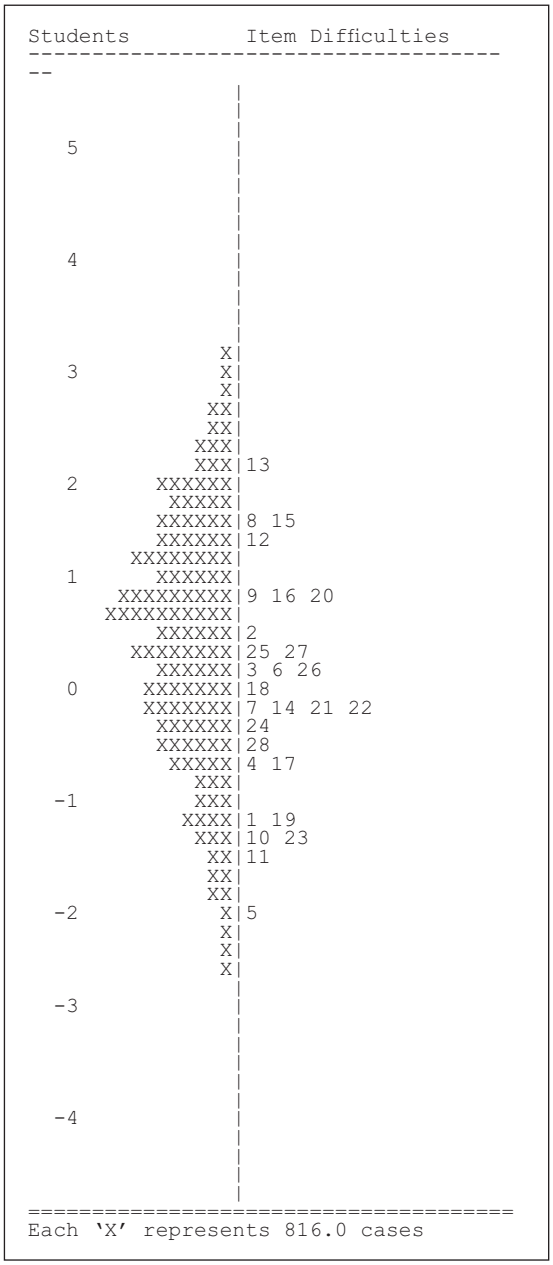


Figure 13.2 ■ **Item plot for reading items**



distribution of student achievement for each of mathematics, reading, science and problem solving, respectively. The figures consist of two panels. The left panel, students, shows the distribution of students' Rasch-scaled achievement estimates. Students at the top end of this distribution have higher achievement estimates than students at the lower end of the distribution. The right panel, item difficulties, shows the distribution of Rasch-estimated item difficulties.

In each of the figures, the student achievement distribution, shown by 'X', is well matched to the item difficulty distribution. The figures are constructed so that when a student and an item are located at the same height on the scale then the student has a 50 per cent chance of responding correctly to the item.

Figure 13.3 ■ **Item plot for science items**

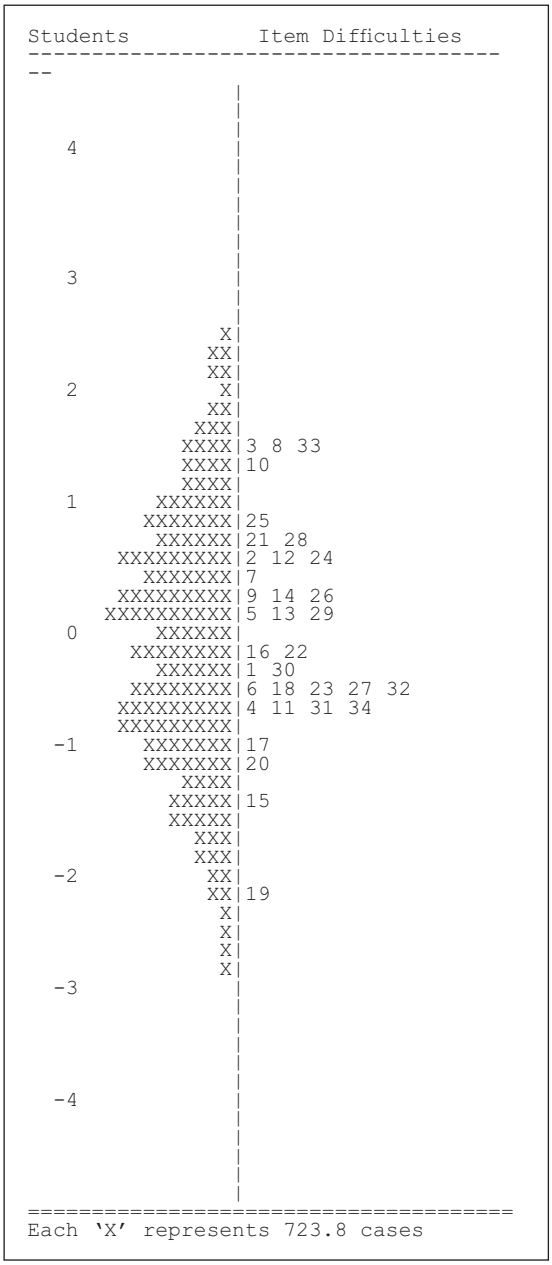
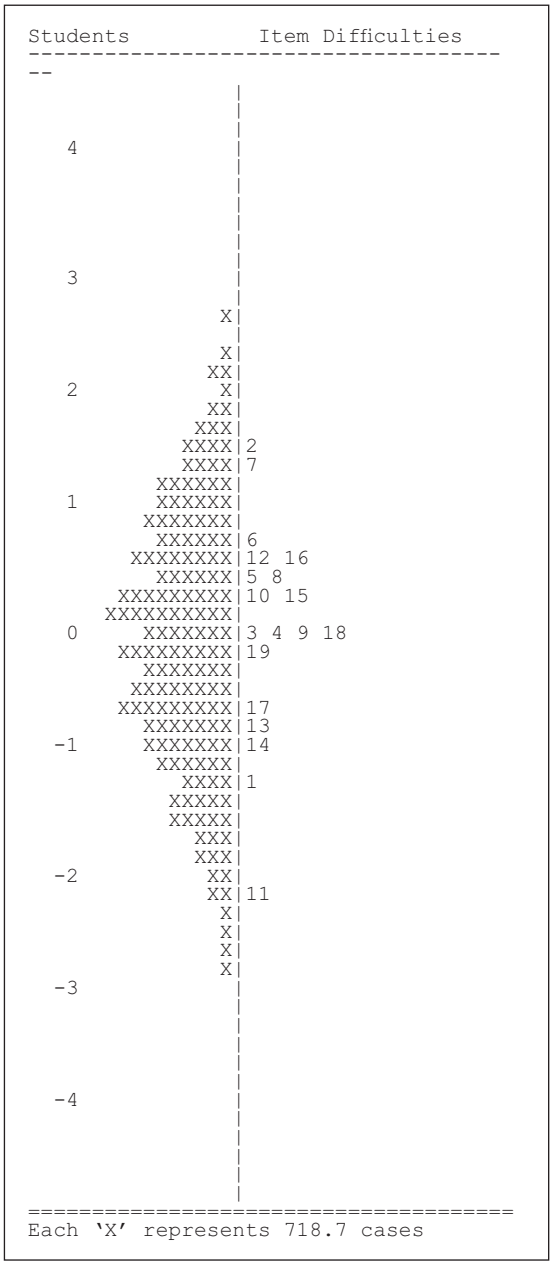


Figure 13.4 ■ **Item plot for problem-solving items**



Test reliability

A second test characteristic that is of importance is the test reliability. Table 13.2 shows the reliability for each of the four overall scales (mathematics, reading, science and problem solving) before conditioning and based upon four separate scalings. The international reliability for each domain after conditioning is reported later in Table 13.6. Appendix 11 shows the reliabilities for each country.

Table 13.2 ■ **Reliabilities of each of the four overall scales when scaled separately**

Domain	Reliability
Mathematics	0.845
Reading	0.799
Science	0.789
Problem solving	0.761

Domain intercorrelations

Correlations between the ability estimates for individual students in each of the four domains, the so-called latent correlations, as estimated by *ConQuest* (Wu *et al.,* 1997) are given in Table 13.3. It is important to note that these latent correlations are unbiased estimates of the true correlation between the underlying latent variables. As such they are not attenuated by the unreliability of the measures, and will generally be higher than the typical product moment correlations that have not been disattenuated for unreliability. The results in the table are reported for both OECD countries and for all participating countries.[1]

Table 13.3 ■ **Latent correlation between the four domains**

	Reading		Science		Problem solving	
	r	SE	r	SE	r	SE
Mathematics						
OECD countries	0.77	0.003	0.82	0.002	0.89	0.001
All participating countries	0.77	0.002	0.82	0.002	0.89	0.001
Reading						
OECD countries			0.83	0.002	0.82	0.002
All participating countries			0.82	0.001	0.82	0.002
Science						
OECD countries					0.79	0.002
All participating countries					0.78	0.002

Mathematics subscales

A seven-dimensional scaling was performed on the achievement data, consisting of:

- Scale 1: mathematics items – space and shape (M1)
- Scale 2: mathematics items – change and relationships (M2)
- Scale 3: mathematics items – uncertainty (M3)
- Scale 4: mathematics items – quantity (M4)
- Scale 5: problem solving items (PS)
- Scale 6: reading items (R)
- Scale 7: science items (S)

Table 13.4 ■ **Correlation between scales**

	M1		M2		M3		M4	
	r	SE	r	SE	r	SE	r	SE
M1								
OECD countries			0.89	0.001	0.88	0.001	0.89	0.001
All participating countries			0.90	0.001	0.89	0.001	0.90	0.001
M2								
OECD countries					0.92	0.001	0.92	0.001
All participating countries					0.92	0.001	0.93	0.001
M3								
OECD countries							0.90	0.001
All participating countries							0.90	0.001
Problem solving								
OECD countries	0.79	0.002	0.83	0.002	0.81	0.002	0.82	0.002
All participating countries	0.80	0.002	0.83	0.001	0.82	0.001	0.83	0.001
Reading								
OECD countries	0.67	0.003	0.73	0.002	0.73	0.002	0.73	0.002
All participating countries	0.68	0.003	0.74	0.002	0.74	0.002	0.73	0.002
Science								
OECD countries	0.73	0.002	0.77	0.002	0.77	0.002	0.76	0.002
All participating countries	0.74	0.002	0.77	0.002	0.78	0.002	0.76	0.002

SCALING OUTCOMES

The procedures for the national and international scaling are outlined in Chapter 9.

Item deletions

The items were first scaled by country and their fit was considered at the national level, as was the consistency of the item parameter estimates across countries. consortium staff then adjudicated items, considering the items' functioning both within and across countries in detail. Those items considered to be "dodgy" (see Chapter 9) were then reviewed in consultation with NPMs. The consultations resulted in the deletion of a few items at the national level and two items at the international level.

At the international level, the two deleted items were S327Q02 and M434Q01T. The nationally deleted items are listed in Table 13.5. All deleted items were recoded as not applicable and were not included in either the international scaling nor in generating plausible values.

Table 13.5 ■ **Items deleted at the national level**

Item	Country
M144Q03	Iceland (booklet 4 only)
M155Q01	Korea
M179Q01T	Italy (Italian version only)
M273Q01	Denmark (booklet 7 only)
M402Q02	Hungary
M442Q02	Uruguay
M603Q02	Canada
M704Q01T	Switzerland (Italian version only)
M800Q01	Uruguay
R055Q03	Austria, Luxembourg (German version only), Germany, Switzerland (German version only), Belgium (German version only), Italy (German version only), Liechtenstein
R102Q04a	Korea
R111Q6B	Tunisia
R219Q01E	Tunisia
R219Q01T	Tunisia
R227Q01	Spain (Catalonian and Castilian versions),
S131Q02T	Russia
S252Q02	Spain (Castilian, Galician, and Valencian versions)
S268Q02T	Norway
S326Q01	Portugal
X414Q01	Russia
X603Q02T	Italy (Italian version only)
X603Q03	Italy (Italian version only)

International scaling

The international scaling was performed on the calibration data set of 15 000 students (500 randomly selected students from each of the 30 OECD

countries). The item parameter estimates from this scaling are reported in Appendices 12, 13, 14 and 15. The item parameters were estimated using four separate one-dimensional models. As discussed later, a booklet facet was used in the item response model.

Generating student scale scores

Applying the conditioning approach described in Chapter 9 and anchoring all of the item parameters at the values obtained from the international scaling, plausible values were generated for all sampled students. Table 13.6 gives the reliabilities at the international level for the generated scale scores. The increase in reliability of the results reported in Table 13.6 over those presented in Table 13.2 is due to the use of multidimensional scaling and conditioning.

Table 13.6 ■ **Final reliability of the PISA scales**

Domain	Reliability
Mathematics (overall)	0.918
Space and shape	0.865
Change and relationships	0.905
Uncertainty	0.905
Quantity	0.895
Reading	0.848
Science	0.843
Problem solving	0.874

TEST LENGTH ANALYSIS

Table 13.7 shows the number of missing responses and the number of missing responses recoded as not reached,[2] by booklet. Table 13.9 shows this information by country.

The average number of not reached items differs from one country to another. It is worth noting that countries with higher averages of not-reached items also have higher averages of missing data. Table 13.8 provides the percentage distribution of not-reached items per booklet. The percentage of students who reached the last item ranges from 77 to 89 per cent (*i.e.* the percentages of students with zero not-reached items).

Table 13.7 ■ **Average number of not-reached items and missing items by booklet**

Booklet	Missing	Not reached
1	4.15	1.34
2	5.43	1.67
3	4.09	0.72
4	5.20	1.19
5	5.93	1.58
6	6.40	1.58
7	5.93	1.52
8	5.52	1.83
9	6.16	2.07
10	5.63	2.07
11	4.80	1.94
12	4.73	2.04
13	5.52	2.19
14	5.10	1.77
Total	**5.34**	**1.67**

Table 13.8 ■ **Percentage distribution of not-reached items by booklet**

Number of not-reached items	Booklet													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
0	84.3	79.0	89.3	81.5	83.9	85.3	84.2	81.4	78.2	77.1	78.2	82.1	77.9	79.3
1	0.6	1.7	2.0	1.2	1.7	0.9	2.0	1.4	3.2	3.2	1.4	0.5	1.7	2.4
2	0.5	3.4	0.9	2.3	1.3	0.8	0.7	2.1	1.9	0.6	1.5	0.9	5.0	0.9
3	0.9	1.6	0.7	2.8	1.0	1.4	1.0	1.5	1.6	1.6	1.0	0.9	0.5	0.9
4	4.5	0.9	0.8	2.1	1.5	0.4	1.0	0.5	0.4	2.6	1.3	2.5	0.6	2.5
5	0.2	2.8	0.4	1.8	1.1	0.7	1.2	1.6	1.0	0.4	4.4	0.4	0.8	1.3
6	0.3	1.1	1.6	1.4	0.7	1.7	1.7	1.0	0.7	1.0	1.1	0.8	0.6	0.3
7	2.1	0.8	0.7	1.6	1.3	0.9	1.6	1.5	0.6	0.8	2.0	1.5	1.6	0.6
8	1.1	1.2	0.8	1.4	0.5	0.7	0.5	1.0	1.3	2.2	1.1	0.3	0.7	2.3
> 8	5.6	7.6	2.9	3.9	7.0	7.2	6.1	8.0	11.1	10.3	8.0	10.3	10.6	9.6

TIMING ISSUES

Timing issues are important for any testing sessions. A test that is too long (ie, contains too many items) will not only frustrate students, but also threaten the validity of the test because many students may rush to complete the test. A test that is too short (ie, too contains too few items) will result in disruptions at the end of the testing session because students will finish the test well before the end of the testing time.

The field trial incorporated procedures to collect some timing information with the consideration that such a collection should not disrupt the testing session, and should not cause undue burden for the students, test administrators and data entry staff.

As a result, five time points were included in each test booklet, requiring students to record the time as they reached these five points. The first time point was at the start of the test questions. The subsequent time points were at the end of each of the four blocks in the test booklet. An example of a timing point is shown below in Figure 13.2.

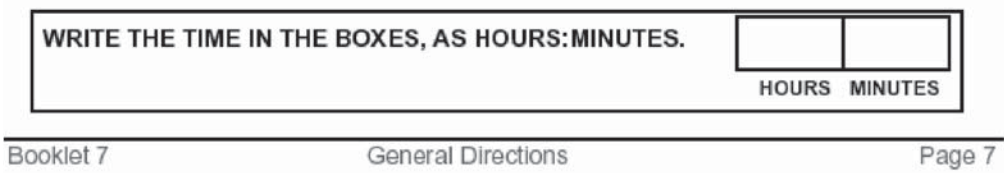
In subsequent sections of this document will denote the five time points as t_1, t_2, t_3, t_4, t_5.

Data collected as described above require a considerable amount of cleaning. Students may move around the test booklet in a somewhat random manner. A simple recording system of time points would not be able to capture such movements, and it was not the intention of this data collection to capture such movements. Consequently, records with any missing time points and any non-increasing time-points were removed from the data analyses. Only timing records with $t_1 < t_2 < t_3 < t_4 < t_5$ were used in the analyses.

Table 13.9. ■ **Average number of missing items and not-reached items by country**

Country	Missing	Not reached
Australia	3.79	0.91
Austria	5.10	0.33
Belgium	4.09	0.99
Brazil	9.32	5.23
Canada	3.04	0.76
Czech Republic	4.62	0.64
Denmark	6.11	1.36
Finland	2.85	0.76
France	5.10	1.32
Germany	5.31	0.62
Greece	8.53	2.45
Hong Kong-China	2.66	0.50
Hungary	5.32	1.43
Iceland	4.11	1.14
Indonesia	7.88	3.43
Ireland	3.30	0.52
Italy	6.07	1.41
Japan	5.24	1.08
Korea	3.46	0.42
Latvia	5.45	1.70
Liechtenstein	3.71	0.40
Luxembourg	6.16	0.89
Macao-China	3.50	1.30
Mexico	6.12	3.73
Netherlands	1.57	0.15
New Zealand	3.55	0.49
Norway	6.65	1.29
Poland	5.99	0.97
Portugal	6.05	1.57
Russian Federation	6.50	3.24
Serbia	11.24	1.47
Slovak Republic	6.10	1.11
Spain	5.76	1.54
Sweden	5.07	1.46
Switzerland	5.11	0.90
Thailand	5.48	2.19
Tunisia	8.81	4.21
Turkey	7.01	1.29
United Kingdom	3.78	0.40
United States	3.23	0.50
Uruguay	10.28	5.56

Figure 13.2 ■ **Example of a timing point**



© OECD 2005

Using these timing markers it is possible to define time intervals $D_{ij} = t_j - t_i$, so that D_{12}, D_{23}, D_{34}, D_{45} are the times taken to complete blocks, 1, 2, 3 and 4 respectively. D_{13} is the time taken to complete the first two blocks and D_{14} is the time taken to complete the first three blocks. Each of the D_{ij} was expressed in hours. In analysing the data, records were deleted for cases in which any of the individual block lengths had a value outside the range of 0 to 1, D_{13} was outside the range 0.5 to 2 and D_{14} was outside the range 0.8 to 2.2. While somewhat arbitrary this choice was made because each block had an expected length of approximately 30 minutes.

There were 28 770 students in the data file, and around 20 000 students had valid timing data. Table 13.10 gives the means and standard deviations of D_{12}, D_{23}, D_{34}, D_{45}, D_{13} and D_{14}, for each booklet.

Table 13.10 ■ **Means, the number of cases and standard deviations of duration time (in hours) by block and by field trial booklet**

Booklet		D12	D23	D34	D45	D13	D14
	Mean	0.5894	0.5469	0.4215	0.3520	1.171	1.553
1	N	1 969	1 958	1 783	1 385	1 958	1 722
	S.D.	0.1736	0.2034	0.1430	0.1313	0.3417	0.3093
	Mean	0.5351	0.5438	0.4719	0.4063	1.105	1.537
2	N	2 039	1 975	1 791	1 211	1 969	1 738
	S.D.	0.1671	0.1976	0.1577	0.1368	0.3340	0.3255
	Mean	0.6006	0.5493	0.4570	0.3496	1.184	1.587
3	N	1 805	1 775	1 553	1 094	1 772	1 480
	S.D.	0.1763	0.2012	0.1473	01392	0.3450	0.3089
	Mean	0.5734	0.5326	0.4819	0.3726	1.37	1.537
4	N	2 140	2 120	1 862	1 298	2 123	1 781
	S.D.	0.1755	0.2040	0.1514	0.1431	0.3494	0.3214
	Mean	0.5965	0.6007	0.4155	0.3122	1.239	1.604
5	N	2 051	2 009	1 819	1 301	2 039	1 753
	S.D.	0.1754	0.2000	0.1495	0.1314	0.3356	0.3022
	Mean	0.5205	0.4149	0.5766	0.3812	0.9631	1.513
6	N	2 125	2 104	1 956	1 428	2 057	1 879
	S.D.	0.1659	0.1685	0.1811	0.1422	0.2935	0.3032
	Mean	0.5308	0.5195	0.3874	0.3914	1.080	1.480
7	N	2 038	2 026	1 961	1 524	2 030	1 873
	S.D.	0.1734	0.11908	0.1385	0.1533	0.3278	0.3128
	Mean	0.5162	0.5813	0.4532	0.3675	1.143	1.560
8	N	1 972	1 895	1 779	1 259	1 918	1 710
	S.D.	0.1824	0.1984	0.1456	0.1493	0.3375	0.3120
	Mean	0.5687	0.4671	0.5076	0.3663	1.070	1.537
9	N	2 050	2 087	1 858	1 327	2 050	1 782
	S.D.	0.1707	0.1852	0.1596	0.1319	0.3302	0.3134
	Mean	0.5481	0.5105	0.4139	0.4201	1.093	1.485
10	N	1 976	1 972	1 858	1 373	1 958	1 787
	S.D.	0.1680	0.1927	0.1451	0.1472	0.3287	0.3162
	Mean	0.5574	0.5253	0.4571	0.3726	1.117	1.535
11	N	20 165	19 921	18 220	13 200	19 874	17 505
	S.D.	0.1754	0.2009	0.1615	0.1439	0.3402	0.3152

Table 13.10 needs to be matched to the test design so that the timing information can be related to the actual test clusters of the assessment material. Table 13.11 shows the PISA 2003 field trial test design.

Table 13.11 ■ **PISA 2003 field trial test design**

Booklet	Block 1 30 minutes	Block 2 30 minutes	Block 3 30 minutes	Block 4 30 minutes
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

For example, the column headed D_{12} in Table 13.10 gives the timing information for mathematics clusters 1 to 10 (M1 to M10), corresponding to the booklets 1 to 10. It can be seen that M8 is the shortest cluster, while M3 is the longest cluster among the first 10 mathematics clusters.

Students were given a break after one hour of testing. The duration of the break varied from country to country. The break was usually just a few minutes in length. The timing information in Table 13.10 includes this break in the computation, as no information is available about the length of the break in this data set. Consequently, the time duration D_{23} is likely to be a slight over-estimate of the actual time taken to complete this block. The over-estimate is probably about 0.05. The fact that the means for the first block are all greater than 0.5 suggests that D_{23} is likely to include this break time, and D_{34} is less likely to include this break. The means for D_{13} are mostly over one hour. That is, the majority of students took more than one hour to complete the first two blocks.

Despite the inclusion of the break time in the computation, there is a trend that the time duration taken to complete a block decreases as testing goes on. For example, the block 3 durations (D_{34}) are all much shorter than block 1 (D_{12}) durations. This observation is consistent with earlier findings that as testing goes on, students' motivation wanes and there are far more missing responses as well as guessing in the latter part of a test than in the earlier part of a test. This observation suggests that the most reliable timing information is the block 1 (D_{12}) information. So these should be used to compute average time taken per item, and not the information from blocks 2, 3 and 4. Certainly, block 4 information is the least reliable. As students run out of time at the end of the test, the block 4 timing cannot be regarded as time taken to complete this block. Nevertheless, blocks 2, 3 and 4 timing information is still useful for comparing the relative lengths of the clusters within these blocks.

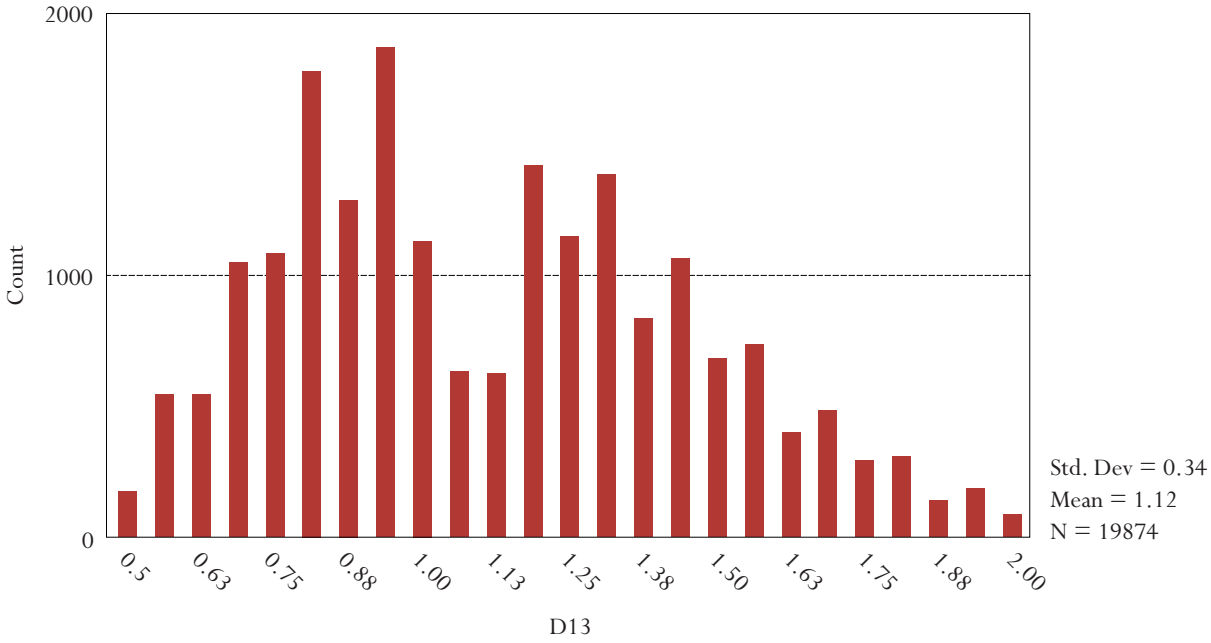
D_{13} provides information about the time taken to complete half of the test. A histogram of D_{13} is shown in Figure 13.3.

The dip in the middle of Figure 13.3 is likely to be caused by the break after one hour of testing. More than half of the students needed more than one hour to complete the first half of the test. Ninety per cent of the students completed the first half of the test after 95 minutes from the start of the test. This suggests that the field trial blocks were, on average, too long (ie, contained too much material) even though 43 per cent completed the first half of the test in less than one hour.

D_{14}, in Table 13.10, provides information about the time taken to complete the first three quarters of the test. A histogram of D_{14} is shown in Figure 13.4.

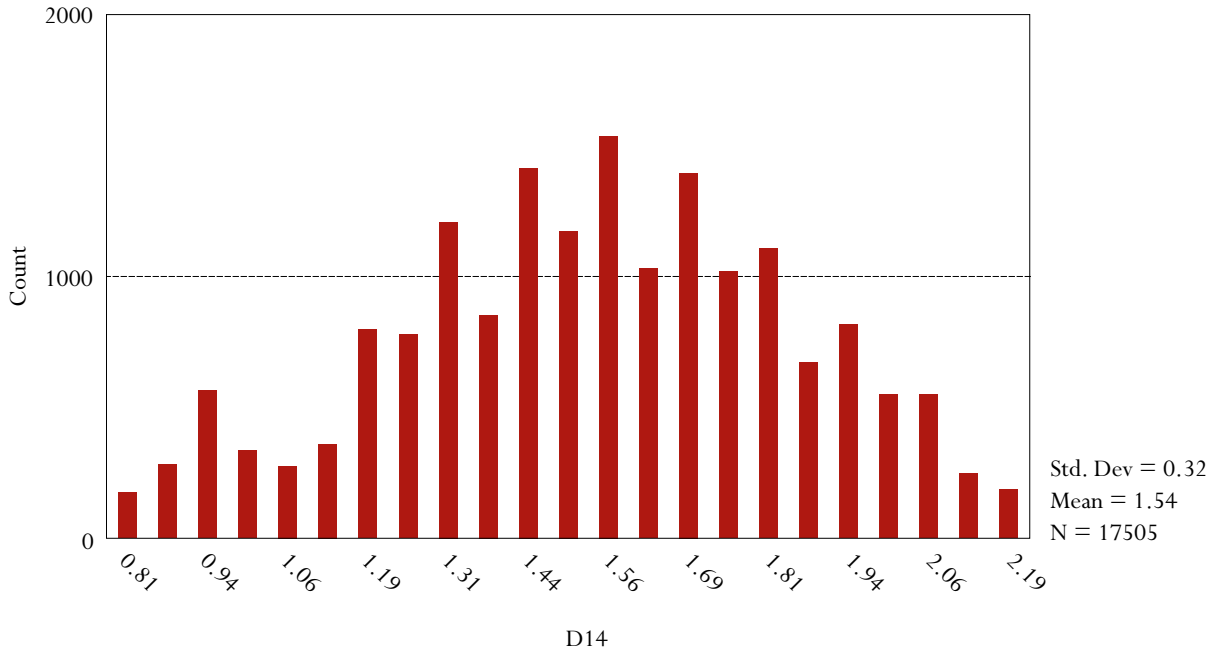
© OECD 2005

Figure 13.3 ■ **Histogram of time taken to complete half of the test**



About half of the students completed the first three quarters of the test one-and-a-half hours after the testing started. And around 92% of students completed the first three-quarters of the test 120 minutes

Figure 13.4 ■ **Histogram of time taken to complete the first three-quarters of the test**



after the testing started. While there is some evidence that the students caught up during the third block, the test still appears to be a little long. Besides, students might have caught up because they knew that time was running out, and started to skip questions or spent less time on some questions.

Average time per item

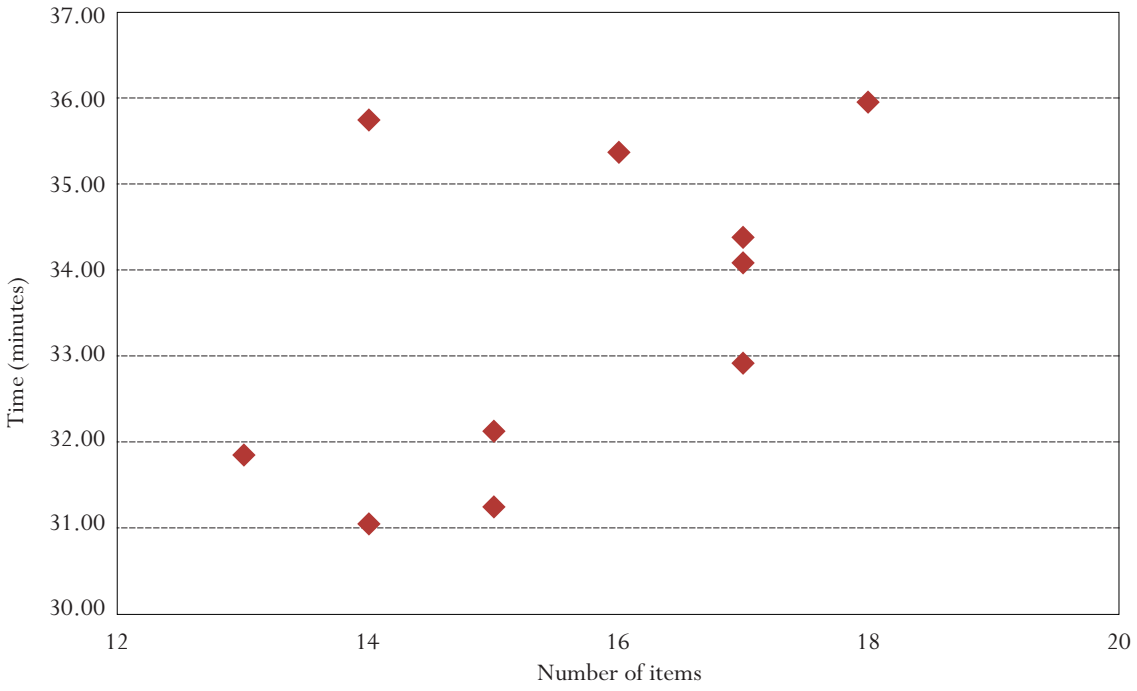
For the first block of the test, Table 13.12 shows the relationship between the number of items, number of words and the average time required to complete the cluster. The average amount of time per item is also reported.

Table 13.12 ■ **Time required in relation to the number of items and number of words**

	Number of items	Number of words	Time (minutes)	
			Per cluster	Per item
M1	16	1338	35.40	2.21
M2	15	1230	32.10	2.14
M3	18	1377	36.00	2.00
M4	17	1277	34.40	2.02
M5	14	1217	35.80	2.56
M6	15	1181	31.20	2.08
M7	13	1061	31.80	2.45
M8	14	1336	31.00	2.21
M9	17	1420	34.10	2.01
M10	17	1294	32.90	1.94
Average			**33.47**	**2.15**

Figure 13.5 shows the relationship between the time required and the number of items in the cluster.

Figure 13.5 ■ **Time required versus the number of items in the cluster**



© OECD 2005

Figure 13.6 ■ **Time required versus the number of words in the cluster**

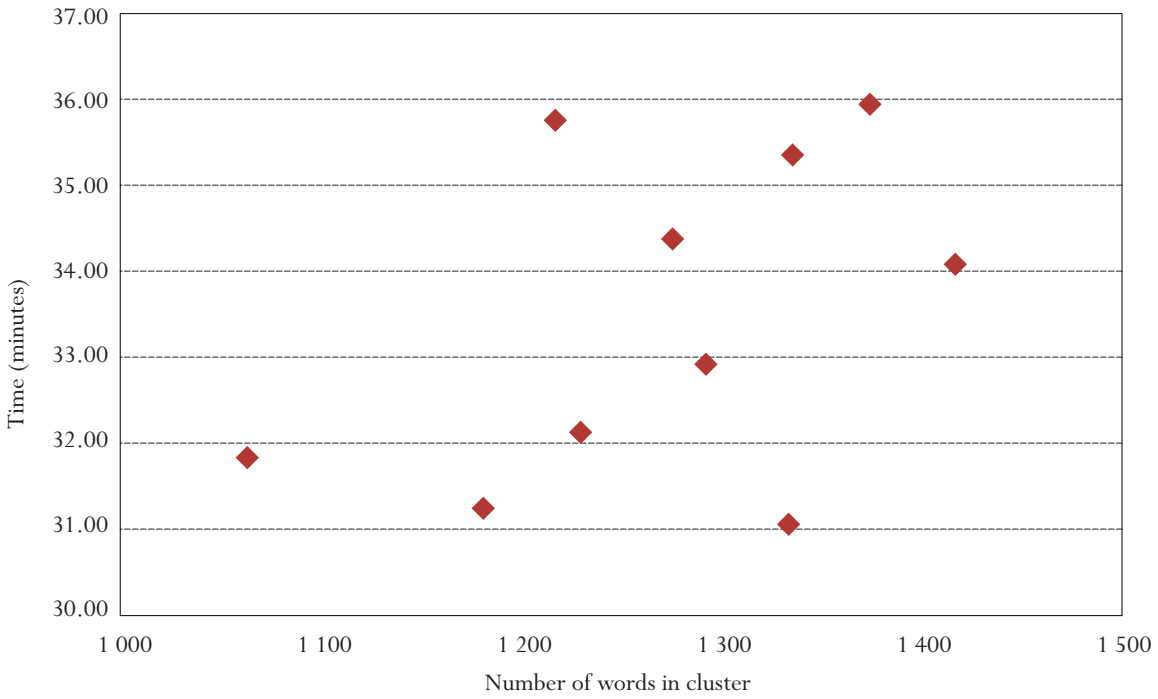


Figure 13.6 shows the relationship between the time required and the number of words in the cluster.

While there is a positive correlation between the time required and the number of items in a cluster, there are also some cases where a cluster with relatively few items took an above average amount of time. For example, a closer examination is required for cluster M5, where the average time per item is the highest. That is, there appear to be other factors (such as item format or item difficulty) that affect the length of time required to complete a cluster. Figure 13.6 shows that there is also a positive correlation between the time required and the number of words in a cluster.

While the average time per item is 2.15 minutes, it should be noted that using this estimate to fill a two-hour test would result in a test that approximately 50 per cent of the students would fail to finish. That is, the time required per item so that around 90 per cent of the students can finish the test within the two-hour time needs to be worked out. For block 1, 90 per cent of the students finished the cluster within 48 minutes. That is an average of 3.08 minutes per item. This is an estimate of the time taken per item when students are considered to be focused on the task. It does not take into account that students will work through the later clusters faster (skipping more items) because of fatigue and loss of motivation.

Number of not-reached items by booklet

At the booklet level, it is also important to monitor the number of not-reached items. The following gives the frequencies of not-reached items for each field trial booklet.

Table 13.13 ■ **Frequencies of not-reached items by field trial booklet**

Booklet	Mean	N	SD
1	3.39	2 879	7.357
2	4.79	2 872	8.609
3	4.94	2 889	8.647
4	3.93	2 891	6.493
5	4.22	2 886	7.616
6	2.57	2 834	4.823
7	1.78	2 860	4.414
8	3.56	2 869	6.600
9	3.58	2 886	6.661
10	2.81	2 874	6.575
97	0.00	11	0.000
99	0.00	19	0.000
Total	356	28 770	6.971

Table 13.13 shows that the average number of not-reached items was between 2 to 5 per booklet. Table 13.13 also suggests that booklets with all mathematics items have higher number of not-reached items. It could be the case that the mathematics clusters are longer on average than problem-solving and science clusters, or that there are more omissions for problem-solving and science items.

Based on the timing distribution for cluster 1, the expected number of not-reached items given a total number of items for a booklet could be computed. Table 13.14 gives the results. Based upon the data in this table a target cluster size of 12 items was adopted for the PISA 2003 main study.

Table 13.14 ■ **The expected number of not-reached items as a function of the total number of items in a booklet**[1]

Assumed time per item (minutes)	Number of items per 30-minute cluster	Number of items in a two-hour booklet	Expected number of not-reached items
1.93	15.6	62	4.4
2.0	15.0	60	3.7
2.1	14.3	57	2.9
2.2	13.6	55	2.3
2.3	13.0	52	1.8
2.4	12.5	50	1.4
2.5	12.0	48	1.1
2.6	11.5	46	0.8
2.7	11.1	44	0.6
2.8	10.7	43	0.5

1. This was the field trial number of items. There were, on average, 15.6 items for the first ten mathematics clusters in the field trial.

BOOKLET EFFECTS

Because the PISA 2003 test design was balanced, the item parameter estimates that are obtained from the scaling are not influenced by a booklet effect, as was the case in PISA 2000. But, due to the different location of domains within each of the booklets it was expected that there would still be booklet influences on the estimated proficiency distributions.

After scaling the PISA 2003 data for each country separately, achievement scores for mathematics, reading, problem solving and science could be compared across countries and across booklets. Tables 13.15, 13.16, 13.17 and 13.18 present student scale scores for the four domains, standardised to have a mean of 10 and

a standard deviation of 2 for each domain and country combination. The table rows represent countries (or sub-regions within countries) and the columns represent booklets. The purpose of these analyses and tables is to examine the nature of any booklet effects, therefore the countries are not named.

If Tables 13.15, 13.16, 13.17 and 13.18 are examined in conjunction with the test design (see Table 2.1 in Chapter 2) the explanation for the patterns in the booklet means is quite clear. From Table 13.15, it can be seen that the mathematics scores are systematically lower on booklets 8, 9, 10 and 11, those being the booklets that only have mathematics at the end. In Table 13.16, the reading scores are systematically lower on booklets 1, 2, 7 and 8. In Table 13.17, the science scores are systematically lower on booklets 5, 6, 12 and 13, and in Table 13.18, the problem-solving scores are systematically lower on booklets 3, 4, 9 and 10.

Table 13.15 ■ **Mathematics means for each country by booklet**

Booklet													
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**UH**
10.40	10.21	10.23	10.16	10.10	10.22	9.94	9.65	9.35	9.38	9.82	10.32	10.19	-
10.28	10.20	10.23	10.31	10.10	9.93	10.07	9.82	9.57	9.61	9.78	9.89	10.20	6.53
10.28	10.38	10.38	10.36	10.18	9.87	10.09	9.40	9.65	9.25	9.92	9.93	10.33	5.68
10.65	10.65	10.78	10.69	10.37	10.17	9.91	9.01	8.46	8.75	9.46	10.49	10.52	-
10.31	10.30	10.22	10.23	10.18	10.13	9.97	9.39	9.51	9.52	9.73	10.26	10.24	-
10.36	10.35	10.31	10.41	10.15	9.95	10.00	9.51	9.57	9.05	9.74	10.24	10.35	-
10.42	10.40	10.25	10.29	10.03	9.90	9.99	9.67	9.50	9.16	9.67	10.30	10.40	7.10
10.21	10.27	10.12	10.16	10.14	10.13	10.17	9.46	9.68	9.31	9.93	10.05	10.39	6.22
10.54	10.36	10.32	10.46	10.14	9.94	9.86	9.22	9.26	9.13	9.94	10.47	10.30	-
10.26	10.50	10.33	10.22	10.01	10.03	10.21	9.07	9.45	9.11	9.99	10.36	10.53	-
10.35	10.57	10.53	10.23	9.98	10.08	10.32	9.37	9.16	8.89	9.64	10.39	10.31	-
10.56	10.40	10.15	10.33	10.16	10.16	10.34	9.27	9.50	9.27	9.54	10.19	10.22	-
10.42	10.69	10.35	10.39	9.99	9.96	10.14	9.21	9.30	9.13	9.59	10.40	10.49	-
10.47	10.32	10.18	10.35	10.12	10.08	9.88	9.51	9.51	9.26	9.88	10.25	10.29	-
10.38	10.56	10.31	10.54	10.32	10.17	10.28	9.02	9.56	9.26	9.53	9.73	10.32	-
10.35	10.18	10.17	10.20	10.14	10.16	9.87	9.54	9.50	9.59	9.93	10.28	10.09	-
10.51	10.53	9.89	10.07	10.33	9.83	10.02	9.17	9.43	9.55	9.72	10.35	10.56	-
10.62	10.71	10.70	10.60	10.22	10.16	10.21	9.02	8.94	8.66	9.24	10.35	10.49	-
10.16	10.02	10.46	10.27	10.00	9.96	10.19	10.16	9.40	9.48	9.76	9.84	10.29	-
10.45	10.40	10.38	10.39	10.27	10.08	10.00	9.37	9.44	9.01	9.66	10.22	10.26	7.97
10.46	10.60	10.84	10.83	10.23	9.59	10.35	9.12	9.18	8.78	9.67	9.62	10.81	-
10.23	10.25	10.32	10.06	10.25	10.30	9.92	9.67	9.25	9.82	9.88	10.02	10.06	-
10.39	10.46	10.44	10.63	10.10	10.21	10.03	9.53	9.37	8.90	9.64	10.13	10.16	-
10.35	10.52	10.23	10.65	10.49	10.18	10.29	8.90	9.31	9.05	9.53	10.53	10.04	-
10.18	10.55	10.50	10.65	10.21	10.10	10.06	9.53	9.11	9.03	9.30	10.27	10.51	-
10.55	10.79	10.25	10.19	10.05	10.24	10.11	9.26	8.95	8.94	9.73	10.49	10.52	-
10.47	10.45	10.51	10.43	10.34	10.10	10.20	9.05	9.08	9.44	9.50	10.14	10.39	-
10.44	10.51	10.17	10.28	10.23	10.05	10.53	9.17	9.50	9.03	9.51	10.22	10.35	-
10.35	10.62	10.29	10.48	10.18	10.16	9.86	9.09	9.59	9.43	10.04	9.74	10.36	-
10.57	10.54	10.63	10.47	10.32	10.18	10.18	8.85	8.95	8.87	9.63	10.47	10.42	-
10.36	10.28	10.39	10.27	10.19	10.10	10.01	10.00	9.27	9.15	9.60	10.11	10.29	-
10.19	10.23	10.28	10.37	10.03	10.00	9.97	10.11	9.57	9.55	9.62	10.08	10.02	-
9.94	10.23	10.39	10.48	10.49	9.96	10.22	9.98	9.36	9.13	9.63	10.04	10.16	-
10.47	10.26	10.46	10.42	10.15	10.13	10.05	9.42	9.25	9.42	9.76	9.94	10.16	-
10.48	10.56	10.60	10.53	10.15	10.06	10.11	9.24	9.28	8.94	9.50	9.87	10.66	-
10.40	10.52	10.88	10.62	9.86	10.00	10.34	9.55	8.80	8.76	9.85	10.15	10.35	-
10.83	10.61	10.76	10.58	10.24	10.11	10.06	8.97	8.47	8.77	9.63	10.26	10.69	-
10.09	10.31	10.12	10.25	10.11	10.02	9.89	9.69	9.67	9.61	9.82	10.09	10.32	5.74
10.30	10.41	10.67	10.27	10.19	10.13	10.08	9.31	9.05	9.23	9.61	10.28	10.42	-
10.28	10.17	10.17	10.09	10.17	10.07	10.12	9.54	9.35	9.45	9.82	10.54	10.19	-
10.49	10.52	10.59	10.53	10.20	9.93	10.10	8.77	9.20	8.92	9.84	10.20	10.73	-
10.49	10.50	10.41	10.34	9.87	10.10	10.27	9.32	9.10	9.33	9.59	10.24	10.39	-
10.49	10.56	10.59	10.71	10.30	10.16	10.17	8.85	8.94	8.81	9.62	9.97	10.72	-
10.32	10.41	10.58	10.40	10.20	9.90	10.15	9.49	9.29	9.09	9.59	10.16	10.43	7.20
10.29	10.46	10.13	10.25	10.09	10.10	10.04	9.08	9.30	9.13	9.94	10.53	10.55	-
10.36	10.57	10.67	10.51	10.21	10.05	10.68	9.02	8.94	9.02	9.70	9.79	10.54	-
10.64	10.83	10.91	10.93	10.25	9.56	10.44	9.03	8.89	8.25	9.72	9.66	10.84	-
10.49	10.44	10.43	10.33	9.96	10.00	10.29	9.37	9.01	9.31	9.88	10.25	10.19	-
10.75	10.89	10.78	10.68	10.23	10.15	10.19	8.55	8.49	8.43	9.61	10.52	10.64	-
10.41	9.59	10.08	9.99	10.34	10.27	10.08	9.44	9.58	9.92	10.05	9.79	10.45	-
10.55	10.55	10.56	10.21	10.38	10.00	10.04	8.96	9.17	8.98	9.60	10.39	10.49	-

Table 13.16 ■ **Reading means for each country by booklet**

Booklet						
1	**2**	**7**	**8**	**9**	**10**	**11**
9.73	9.61	9.65	9.96	10.26	10.31	10.50
9.78	9.85	9.87	10.03	9.97	10.22	10.28
9.89	9.57	9.67	9.82	10.12	10.38	10.54
9.53	8.71	10.07	9.26	10.87	10.92	10.78
9.80	9.86	9.50	9.74	10.16	10.38	10.56
9.76	9.73	9.63	10.01	10.00	10.24	10.63
9.81	9.71	9.82	9.79	10.16	10.26	10.45
10.04	9.85	9.81	9.81	9.98	10.26	10.25
9.93	9.70	9.37	9.80	9.90	10.44	10.93
9.36	9.40	9.62	9.86	10.30	10.52	10.86
9.68	9.59	9.64	9.90	10.40	10.36	10.51
9.25	9.34	9.60	9.85	10.17	10.99	10.73
9.61	9.64	9.51	9.69	10.37	10.32	10.82
9.84	9.55	9.75	9.72	10.38	10.29	10.44
9.69	9.69	9.82	9.86	10.05	10.32	10.58
9.99	9.85	9.64	9.70	10.11	10.20	10.53
9.90	10.12	9.44	9.64	9.80	10.46	10.69
9.32	9.10	9.65	9.69	10.56	10.68	11.01
9.77	10.03	9.50	9.92	9.85	10.25	10.70
9.67	9.46	9.67	9.92	10.32	10.33	10.68
9.41	9.12	9.71	9.63	10.44	10.64	10.98
9.92	10.01	9.64	9.91	9.93	10.15	10.44
9.76	9.65	9.69	9.99	10.13	10.40	10.39
10.01	9.67	9.54	9.64	10.27	10.46	10.35
10.03	9.61	9.89	9.75	10.05	10.41	10.23
9.67	9.27	9.63	9.89	10.32	10.55	10.67
9.76	9.41	9.67	9.68	10.16	10.71	10.55
9.56	9.44	9.77	10.08	10.24	10.49	10.42
9.51	9.38	9.76	10.00	10.40	10.16	10.67
9.54	9.36	9.51	9.65	10.42	10.63	10.88
9.65	9.93	9.52	10.00	9.79	10.34	10.77
9.62	9.92	9.22	9.96	9.90	10.58	10.78
9.85	9.76	9.79	10.44	9.82	10.19	10.13
9.84	9.58	9.75	9.73	10.29	10.29	10.58
9.49	9.52	9.66	10.05	10.39	10.35	10.57
9.50	9.60	9.58	9.85	10.04	10.29	11.13
9.50	8.97	9.73	9.68	10.47	10.75	10.90
9.61	9.91	9.52	10.06	9.95	10.26	10.69
9.50	9.80	9.71	9.57	10.25	10.49	10.67
9.93	9.84	9.77	9.75	10.01	10.17	10.52
9.95	9.29	9.90	9.75	10.28	10.33	10.53
9.95	9.16	9.98	9.75	10.46	10.41	10.34
9.20	9.21	9.77	9.94	10.38	10.72	10.83
9.64	9.65	9.64	9.99	10.11	10.33	10.64
9.48	9.75	9.63	9.85	10.27	10.46	10.64
9.48	9.33	9.59	9.63	10.32	10.64	11.00
9.32	9.00	9.61	9.56	10.60	10.68	11.22
9.68	9.45	9.63	9.76	10.20	10.40	10.95
9.17	9.12	9.65	9.67	10.70	10.68	11.12
9.94	9.63	9.90	9.53	10.16	10.31	10.52
9.75	9.47	9.70	9.50	10.25	10.54	10.81

Table 13.17 ■ **Science means for each country by booklet**

Booklet						
5	**6**	**7**	**8**	**9**	**12**	**13**
9.87	9.96	10.26	10.30	10.41	9.50	9.70
9.77	10.18	9.94	10.25	10.54	9.38	9.95
10.11	9.93	10.26	10.31	10.41	9.29	9.71
9.95	10.05	10.55	10.73	10.82	8.77	9.18
9.95	10.07	10.16	10.28	10.26	9.41	9.88
9.92	10.08	10.09	10.35	10.41	9.34	9.81
9.80	10.04	10.04	10.22	10.58	9.32	9.99
9.89	10.14	9.93	10.22	10.40	9.34	10.06
10.13	9.67	10.33	10.31	10.20	9.73	9.67
9.68	10.33	10.23	10.47	10.83	8.85	9.57
9.88	10.30	10.24	10.30	10.71	8.83	9.78
10.02	9.84	10.09	10.39	11.23	8.94	9.51
9.99	9.94	10.31	10.45	10.71	9.24	9.33
9.81	10.14	10.19	10.29	10.52	9.21	9.78
10.27	9.80	10.48	10.37	10.32	9.47	9.31
9.96	9.96	10.06	10.21	10.48	9.60	9.73
10.38	9.64	10.19	9.98	10.08	9.86	9.87
9.59	10.25	10.23	10.79	11.13	8.72	9.34
9.77	10.16	9.89	10.11	10.44	9.55	10.10
9.76	10.52	9.86	10.41	10.89	8.91	9.65
10.15	9.77	10.80	10.63	10.69	9.11	8.87
9.85	10.05	10.03	10.41	10.43	9.61	9.61
9.83	10.20	10.16	10.50	10.63	8.91	9.73
9.79	10.26	10.21	10.43	10.72	8.98	9.60
9.47	10.61	10.31	9.91	10.77	9.18	9.77
9.53	10.13	10.18	10.63	10.82	8.93	9.77
9.81	10.30	10.07	10.47	10.71	9.00	9.73
9.90	10.02	10.19	10.61	10.71	8.96	9.58
9.82	10.25	9.97	10.39	10.66	8.85	9.95
9.58	10.35	10.27	10.49	10.95	8.95	9.38
9.69	10.45	9.71	10.32	10.93	8.94	9.97
9.84	10.09	10.06	10.33	10.41	9.56	9.69
9.82	9.92	10.35	10.78	10.24	9.26	9.75
9.90	10.12	10.03	10.28	10.64	9.27	9.73
9.71	10.28	10.10	10.60	10.73	8.98	9.60
9.67	10.47	10.17	10.47	10.78	9.05	9.31
9.61	10.52	10.16	10.69	11.25	8.44	9.30
10.11	9.84	10.08	10.22	10.34	9.66	9.74
9.94	10.10	10.31	10.19	10.58	9.33	9.59
9.93	9.90	10.28	10.07	10.41	9.60	9.86
9.89	10.22	9.93	10.35	10.72	9.31	9.62
9.85	10.00	10.33	10.45	10.49	9.28	9.63
9.79	10.33	10.29	10.62	10.83	8.77	9.41
9.76	10.16	10.04	10.46	10.66	9.08	9.82
9.98	9.99	10.27	10.33	10.47	9.41	9.56
9.98	9.80	10.50	10.40	10.86	9.20	9.19
9.97	9.67	10.73	10.81	10.87	8.73	9.23
9.97	10.00	10.28	10.38	10.60	9.21	9.52
9.91	10.04	10.57	10.79	11.02	8.72	8.91
9.89	10.07	10.16	10.30	10.53	9.51	9.53
9.90	10.29	10.12	10.37	10.80	8.96	9.54

Table 13.18 ■ **Problem-solving means for each country by booklet**

Booklet						
3	**4**	**9**	**10**	**11**	**12**	**13**
9.68	9.65	9.86	9.98	10.15	10.17	10.50
9.92	9.96	9.81	9.99	10.09	10.07	10.15
9.77	9.59	9.88	10.02	10.30	10.11	10.31
9.74	8.81	10.06	9.61	10.73	10.53	10.51
9.66	9.75	9.81	10.05	9.68	10.37	10.66
9.68	9.75	9.68	10.04	10.02	10.25	10.57
9.67	9.73	9.83	9.90	10.16	10.25	10.44
9.72	9.91	9.76	9.93	10.16	10.21	10.31
9.73	9.60	9.86	9.90	10.23	10.37	10.34
9.52	9.37	9.88	9.99	10.32	10.45	10.47
9.79	9.27	10.20	9.86	9.96	10.32	10.60
9.38	9.34	9.92	10.35	10.09	10.20	10.60
9.55	9.21	10.06	10.00	10.33	10.38	10.51
9.77	9.35	9.94	10.11	10.24	10.14	10.43
9.73	9.60	9.89	9.88	10.06	10.29	10.54
9.69	9.92	9.77	10.11	10.06	10.13	10.32
9.32	9.67	9.65	10.16	10.14	10.25	10.84
9.50	9.00	9.78	9.68	10.45	10.74	10.80
10.24	9.92	9.87	9.60	10.26	10.01	10.11
9.74	9.22	10.01	9.89	10.37	10.51	10.25
9.94	8.78	10.08	9.49	10.74	10.72	10.25
9.96	9.29	10.01	9.93	10.30	10.26	10.23
9.62	9.44	10.00	9.88	10.08	10.47	10.52
9.32	9.47	9.72	10.05	10.35	10.67	10.44
9.13	9.42	9.73	10.27	10.12	10.37	10.95
9.34	9.09	9.90	9.91	10.41	10.72	10.62
9.59	9.31	9.89	9.89	10.35	10.34	10.59
9.32	9.22	10.10	10.35	10.23	10.38	10.45
9.38	10.09	10.03	10.06	10.49	9.85	10.08
9.77	9.07	9.84	9.68	10.52	10.45	10.63
10.08	9.67	9.90	9.49	10.25	10.24	10.38
10.11	9.65	10.05	9.76	10.07	10.19	10.16
10.12	9.97	9.43	10.05	9.90	10.02	10.44
9.80	9.70	9.74	9.92	10.14	10.17	10.56
9.79	9.17	10.01	9.60	10.41	10.41	10.62
10.29	9.24	9.82	9.34	10.53	10.54	10.30
9.53	8.80	9.99	9.54	10.90	10.63	10.64
9.53	9.75	9.65	10.10	10.15	10.23	10.58
10.01	9.30	9.84	9.72	10.24	10.34	10.55
9.61	9.85	9.72	9.95	9.97	10.11	10.75
9.70	9.56	9.95	9.75	10.16	10.37	10.50
9.59	9.17	9.94	9.92	10.26	10.37	10.74
9.33	9.00	9.97	9.88	10.53	10.50	10.86
9.97	9.38	9.92	9.78	10.27	10.34	10.36
9.53	9.32	9.89	9.97	10.36	10.48	10.52
10.00	9.09	10.13	9.27	10.83	10.68	10.05
9.41	9.02	9.67	9.54	10.49	10.86	10.98
9.90	9.43	9.65	9.90	10.41	10.30	10.43
9.43	8.67	9.81	9.67	10.65	10.87	10.99
9.67	9.69	9.77	10.21	10.04	10.17	10.44
9.83	9.27	9.82	9.69	10.31	10.57	10.52

Correction of the booklet effect

Modelling the effect

Modelling the order effects in terms of item positions in a booklet or at least in terms of cluster positions in a booklet would result in a very complex model. For the sake of simplicity in the international scaling, the effect, as in PISA 2000, was modelled at the booklet level, separately for each domain.

When estimating the item parameters, booklet effects were included in the measurement model to prevent confounding item difficulties and booklet effects. For the *ConQuest* model statement, the calibration model was: item + item*step + booklet.

The booklet parameter, formally defined in the same way as item parameters, reflects booklet difficulty.[3]

Estimating the parameters

The calibration model given above was used to estimate the international item parameters. It was estimated using the international calibration sample of 15 000 students, and not-reached items in the estimation were treated as not administered.

The booklet parameters obtained from this analysis were not used to correct for the booklet effect. Instead, a set of booklet parameters was obtained by scaling the entire data set of OECD countries using booklet as a conditioning variable and a senate weight. The students who responded to the UH booklet were excluded from the estimation. The booklet parameter estimates obtained are reported in Table 13.19. The booklet effects are the amount that must be added to the proficiencies of student who responded to each booklet. That is a positive value indicates a booklet that was harder than the average while a negative value indicates a booklet that was easier than the average. Since the booklet effects are deviations from an average they sum to zero for each domain.

Table 13.20 shows the booklet effects after transformation to the PISA scales.

Table 13.19 ■ **Estimated booklet effects in logits**

	Domain			
Booklet	**Mathematics**	**Reading**	**Science**	**Problem solving**
1	–0.24	0.18		
2	-0.22	0.24		
3	-0.21			0.16
4	-0.20			0.27
5	-0.09		0.07	
6	-0.05		-0.06	
7	-0.04	0.20	-0.09	
8	0.36	0.11	-0.20	
9	0.41	-0.12	-0.33	0.07
10	0.46	-0.23		0.06
11	0.15	-0.38		-0.13
12	-0.13		0.41	-0.17
13	-0.21		0.19	-0.26

Table 13.20 ■ **Estimated booklet effects on the PISA scale**

	Domain			
Booklet	**Mathematics**	**Reading**	**Science**	**Problem solving**
1	-18.5	14.0		
2	-17.1	19.3		
3	-16.4			13.5
4	-15.5			23.2
5	-6.8		6.4	
6	-3.7		-5.5	
7	-2.8	16.1	-7.8	
8	27.9	8.7	-18.0	
9	31.5	-9.4	-29.7	6.0
10	35.7	-18.1		4.9
11	12.0	-30.6		-11.2
12	-10.2		37.2	-14.5
13	-16.1		17.3	-21.9

Applying the correction

To correct the student scores for the booklet effects, two alternatives were considered:

- Correcting all students' scores using one set of the internationally estimated booklet parameters; or
- Correcting the students' scores using nationally estimated booklet parameters for each country.

When choosing between these two alternatives a number of issues were considered. First, it is important to recognise that the sum of the booklet correction values is zero for each domain, so the application of either of the above corrections does not change the country means or rankings. Second, if a national correction was applied then the mean, within a country, will be the same for each booklet. As such, this approach would incorrectly remove a component of expected sampling and measurement error variation. Third, the booklet corrections are essentially an additional set of item parameters that capture the effect of the item locations in the booklets.

In PISA all item parameters are treated as international values so that all countries are therefore treated in exactly the same way. Perhaps the following scenario best illustrates the justification for this. Suppose students in a particular country found the reading items on a particular booklet surprisingly difficult, even though those items have been deemed as central to the PISA definition of PISA reading literacy and have no technical flaws, such as a translation or coding error. If a national correction were used, then an adjustment would be made to compensate for the greater difficulty of these items in that particular country. The outcome would be that two students from two different countries who responded in the same way to these items would be given different proficiency estimates. This differential treatment of students based upon their country has not been deemed as suitable in PISA. Moreover, this form of adjustment would have the effect of masking real underlying differences in literacy between students in those two countries, as indicated by those items.

Applying an international correction was therefore deemed the most desirable option from the perspective of cross-national consistency.

Table 13.21 ■ Standard deviations of mean scores across booklets

	Mathematics		Reading		Science		Problem solving	
	SD of booklet means	SE of PISA mean	SD of booklet means	SE of PISA mean	SD of booklet means	SE of PISA mean	SD of booklet means	SE of PISA mean
Australia	6.24	2.15	4.71	2.13	7.09	2.10	5.59	1.98
Austria	10.85	3.27	11.31	3.76	9.05	3.44	12.38	3.18
Belgium	6.52	2.29	4.98	2.58	8.24	2.48	5.44	2.20
Brazil	23.00	4.83	36.29	4.58	17.08	4.35	19.67	4.84
Canada	6.22	1.82	7.03	1.75	9.04	2.02	11.95	1.74
Czech Republic	5.63	3.55	5.24	3.46	7.56	3.38	5.97	3.42
Denmark	7.22	2.74	10.02	2.82	16.49	2.97	4.27	2.54
Finland	5.67	1.87	9.30	1.64	5.18	1.92	5.82	1.86
France	11.00	2.50	5.14	2.68	18.17	2.99	4.13	2.67
Germany	7.93	3.32	11.49	3.39	10.34	3.64	9.14	3.24
Greece	17.64	3.90	22.50	4.10	22.20	3.82	19.45	3.97
Hong Kong-China	13.59	4.54	12.26	3.69	14.21	4.26	18.25	4.18
Hungary	4.32	2.84	3.85	2.47	13.68	2.77	8.90	2.86
Iceland	7.31	1.42	6.00	1.56	7.87	1.47	4.82	1.38
Indonesia	15.21	3.91	8.90	3.38	15.02	3.21	15.59	3.29
Ireland	12.32	2.45	11.81	2.63	9.01	2.69	8.48	2.34
Italy	10.63	3.08	9.09	3.04	13.96	3.13	11.20	3.10
Japan	9.18	4.02	12.63	3.92	20.30	4.14	12.43	4.05
Korea	12.54	3.24	11.94	3.09	7.96	3.54	11.55	3.06
Latvia	11.37	3.69	6.60	3.67	5.27	3.89	11.54	3.90
Liechtenstein	12.87	4.12	18.31	3.58	12.89	4.33	17.16	3.95
Luxembourg	5.80	0.97	4.34	1.48	3.64	1.50	6.14	1.37
Macao-China	13.87	2.89	6.11	2.16	12.85	3.03	16.46	2.53
Mexico	15.43	3.64	13.21	4.09	19.03	3.49	19.65	4.30
Netherlands	10.34	3.13	9.78	2.85	12.42	3.15	8.48	2.95
New Zealand	7.75	2.26	8.25	2.46	11.55	2.35	11.84	2.17
Norway	7.04	2.38	9.75	2.78	6.05	2.87	9.12	2.60
Poland	12.32	2.50	9.24	2.88	7.50	2.86	3.60	2.78
Portugal	6.73	3.40	13.42	3.73	5.04	3.46	9.37	3.87
Russian Federation	15.63	4.20	17.15	3.94	16.61	4.14	19.88	4.59
Serbia	9.22	3.75	6.61	3.56	5.24	3.50	7.59	3.32
Slovak Republic	5.89	3.35	5.24	3.12	6.48	3.71	5.52	3.38
Spain	6.01	2.41	7.95	2.60	11.08	2.61	5.74	2.73
Sweden	9.18	2.56	6.25	2.42	7.33	2.72	6.28	2.44
Switzerland	4.68	3.38	8.30	3.28	7.64	3.69	7.53	3.05
Thailand	12.70	3.00	8.62	2.81	11.62	2.70	21.10	2.72
Tunisia	21.83	2.54	23.19	2.81	20.05	2.56	16.78	2.11
Turkey	10.13	6.74	5.99	5.79	5.46	5.89	8.38	6.03
United Kingdom	9.36	2.43	8.11	2.46	10.58	2.52	11.63	2.38
United States	17.58	2.95	8.65	3.22	7.58	3.08	8.07	3.13
Uruguay	31.35	3.29	34.75	3.43	33.51	2.90	35.52	3.68

Remaining booklet effects

The choice of a common correction does, however, leave deviations from equal booklet means in the data and these deviations vary over countries. These deviations occur because of sampling error, measurement error and any remaining item- or booklet-by-country interactions in the data. The results in Appendix 3 show the mean for each country by booklet after the international correction has been implemented. The annexes also show the country ranks that would have resulted using each booklet.

In Table 3.21, the results in the appendix are summarised by showing the standard deviation of the means across booklets. As a point of comparison the standard error of the PISA mean is also shown.

Under the assumption that the scaling model is correct, all of the variation between the booklet means should be explainable through sampling and measurement error. While there is variation across countries and booklet in the standard errors of the booklet means, they are typically about two to three times the size of the standard error of the PISA mean. It follows that where the standard deviations of the booklet means exceed the standard error of the PISA means by a factor of about three, there are remaining item- or booklet-by-country interactions in the data. The observation of these booklet variations is an important outcome of PISA that should not be neglected when analysing, reporting and interpreting PISA results.

Imputing data for students who did not respond to a domain

The PISA conditioning variables are prepared using procedures based on those used in the United States National Assessment of Educational Progress (Beaton, 1987) and in TIMSS (Macaskill *et al.,* 1998). The steps involved in this process are as follows:

- *Step 1.* Five variables (booklet ID, gender, mother's occupation, father's occupations and school mean mathematics score) were prepared to be directly used as conditioning variables. The booklet ID was dummy coded so that booklet 9 was used as the reference booklet. Booklet 9 had to be chosen as the reference booklet because it is the only booklet that contains items from all four assessment domains. For mother's and father's occupation the international socio-economic index of occupational status (ISEI) was used. For each student the mean mathematics achievement for that student's school was estimated using the mean of the weighted likelihood estimates for mathematics for each of the students who also attended that student's school.
- *Step 2.* Each variable in the Student Questionnaire was dummy coded. The details of this dummy coding are provided in Appendix 10.
- *Step 3.* For each country, a principal components analysis of the dummy-coded variables was performed, and component scores were produced for each student (a sufficient number of components to account for 95 per cent of the variance in the original variables).
- *Step 4.* The item-response model was fitted to each national data set and the national population parameters were estimated using item parameters anchored at their international location, and conditioning variables derived from the national principal components analysis and from *Step 1*.
- *Step 5.* Five vectors of plausible values were drawn using the method described in Chapter 9. The vectors were of length seven, one for each of the PISA 2003 reporting scales.

In PISA 2000 the plausible values for those students who did not respond to any items from a domain were removed from the database and a set of weight adjustments was provided for dealing with the smaller data

set. The assumption under this approach is that the students who did not get domain scores were missing at random. For PISA 2003 the plausible values for all domains have been retained for all students. This approach has a number of advantages. First, the database structure is simpler and analysis is simpler because the use of a weight adjustment is not necessary. Second, the missing at random assumption is loosened somewhat. The plausible value generation assumes that the relationships between the domain for which no items are observed and all other variables (both conditioning variables and the other domains) is the same for both the students who did respond to items from a domain and those who did not. Using all of this relationship information, and all available information about the student, an imputation is made. Because of the amount of data that is available to make the imputation, the analysis of the full data set will produce more accurate results than will analysis of the data set that omits students who did not respond to a domain. Additionally, it can be expected that, due to sampling variation, the characteristics of the students who did not respond to a domain will be slightly different from the characteristics of those who did, these differences will be appropriately adjusted for in the imputation and the estimated characteristics of, for example, the reading proficiency distribution for all students will be slightly different from the estimated characteristics of the reading proficiency distribution for the subset of students who responded to reading items.

The one disadvantage of this approach is that the average performances on the reference booklet (booklet 9) will influence the imputations for students who did not respond to items from a domain. As noted above, booklet- and item-by-country interactions do result in variations across booklets in the country means. If a country has an unusually low or high performance on the reference booklet, for a particular domain, then this unusual performance will influence the imputations for all students that did not respond to that domain. The consequential effect is that the reference booklet will be given more weight than the other booklets in the assessment of national means.

Tables 13.22, 13.23 and 13.24 show the mean and standard errors of the mean for each country using all students in the database, and using the subset of students who responded to items in each domain for reading, science and problem solving. The tables also show the difference between the mean of all students and the mean of the assessed students and the ratio of the error variances for the two estimates of the mean.

For the majority of the cases the variance ratio is less than one. This indicates that the error variances associated with the estimate of the mean for all students is less than that for the assessed students. It is important to realise that this is not an artificial result that is merely due to an increase in sample size, but is a genuine reduction in the error caused by the increase in the total available information about the proficiency distribution.

For a number of countries the difference between the means is reasonably large. In the case of reading, amongst OECD countries the difference is significant for Denmark. For science the differences are significant for the following OECD countries: Canada, Denmark, Greece and Mexico. For problem solving, none of the differences are significant for OECD countries.

Table 13.22 ■ Comparison of reading means for all students and reading-assessed students

	All students		Reading-assessed students only		Difference	Ratio of error variance
	Mean	SE of mean	Mean	SE of mean	(All - Assessed)	(All/Assessed)
Australia	525	2.1	526	2.1	-0.8	1.01
Austria	491	3.8	499	4.0	-7.8	0.89
Belgium	507	2.6	508	2.9	-1.3	0.78
Brazil	403	4.6	383	5.3	19.6	0.76
Canada	528	1.7	529	1.9	-1.3	0.87
Czech Republic	489	3.5	490	3.6	-1.6	0.93
Denmark	492	2.8	502	3.2	-9.3	0.79
Finland	543	1.6	541	2.1	2.6	0.62
France	496	2.7	499	3.1	-2.5	0.73
Germany	491	3.4	493	3.7	-2.1	0.86
Greece	472	4.1	460	4.4	12.4	0.87
Hong Kong-China	510	3.7	517	4.4	-7.1	0.72
Hungary	482	2.5	480	2.9	1.5	0.73
Iceland	492	1.6	492	2.1	-0.3	0.53
Indonesia	382	3.4	379	3.5	3.1	0.94
Ireland	515	2.6	521	3.1	-5.6	0.72
Italy	476	3.0	471	3.4	4.2	0.82
Japan	498	3.9	507	3.9	-9.0	0.99
Korea	534	3.1	540	3.3	-5.8	0.89
Latvia	491	3.7	486	3.9	4.5	0.89
Liechtenstein	525	3.6	528	5.8	-3.3	0.38
Luxembourg	479	1.5	479	1.9	0.5	0.61
Macao-China	498	2.2	500	3.3	-2.2	0.43
Mexico	400	4.1	394	4.5	5.5	0.82
Netherlands	513	2.9	517	3.0	-3.6	0.91
New Zealand	522	2.5	524	2.8	-2.0	0.75
Norway	500	2.8	495	3.1	4.3	0.81
Poland	497	2.9	494	3.1	2.8	0.87
Portugal	478	3.7	473	3.9	4.7	0.93
Russian Federation	442	3.9	439	4.4	3.4	0.79
Serbia	412	3.6	412	3.8	0.0	0.88
Slovak Republic	469	3.1	470	3.3	-0.9	0.89
Spain	481	2.6	478	3.1	2.3	0.71
Sweden	514	2.4	515	2.9	-0.7	0.70
Switzerland	499	3.3	504	3.8	-5.1	0.75
Thailand	420	2.8	419	2.8	1.0	1.02
Tunisia	375	2.8	367	2.9	7.6	0.93
Turkey	441	5.8	439	6.0	2.0	0.92
United Kingdom	507	2.5	508	2.9	-1.2	0.74
United States	495	3.2	495	3.7	-0.3	0.75
Uruguay	434	3.4	417	3.9	17.2	0.77

Table 13.23 ■ Comparison of science means for all students and science assessed students

	All students		Science assessed students only		Difference	Ratio of error variance
	Mean	SE of mean	Mean	SE of mean	(All – Assessed)	(Assessed/All)
Australia	525	2.1	531	2.3	-5.8	0.80
Austria	491	3.4	495	3.4	-3.8	1.00
Belgium	509	2.5	512	2.6	-3.5	0.90
Brazil	390	4.3	386	4.4	3.2	0.98
Canada	519	2.0	528	2.3	-9.3	0.76
Czech Republic	523	3.4	523	3.9	0.0	0.74
Denmark	475	3.0	486	3.1	-11.1	0.94
Finland	548	1.9	551	2.3	-2.5	0.72
France	511	3.0	516	2.9	-4.8	1.06
Germany	502	3.6	506	3.7	-3.4	0.95
Greece	481	3.8	465	3.9	16.0	0.94
Hong Kong-China	539	4.3	545	4.5	-5.1	0.90
Hungary	503	2.8	498	3.0	5.0	0.85
Iceland	495	1.5	493	2.1	2.0	0.49
Indonesia	395	3.2	398	3.5	-2.7	0.83
Ireland	505	2.7	511	3.0	-5.3	0.79
Italy	486	3.1	480	3.4	6.9	0.83
Japan	548	4.1	536	4.6	11.2	0.82
Korea	538	3.5	541	3.8	-2.4	0.87
Latvia	489	3.9	487	4.5	1.7	0.74
Liechtenstein	525	4.3	532	6.9	-6.5	0.39
Luxembourg	483	1.5	482	2.0	1.1	0.56
Macao-China	525	3.0	517	4.1	7.7	0.56
Mexico	405	3.5	393	3.9	12.2	0.80
Netherlands	524	3.1	529	3.5	-4.2	0.81
New Zealand	521	2.4	525	2.5	-3.7	0.86
Norway	484	2.9	483	3.6	1.3	0.62
Poland	498	2.9	493	3.5	4.3	0.68
Portugal	468	3.5	472	3.8	-4.4	0.85
Russian Federation	489	4.1	485	4.4	4.3	0.87
Serbia	436	3.5	434	3.5	2.0	0.97
Slovak Republic	495	3.7	493	4.1	2.2	0.83
Spain	487	2.6	480	2.9	6.8	0.79
Sweden	506	2.7	510	2.8	-3.8	0.97
Switzerland	513	3.7	517	3.9	-3.9	0.90
Thailand	429	2.7	425	3.0	4.2	0.79
Tunisia	385	2.6	380	2.7	4.5	0.89
Turkey	434	5.9	433	6.0	1.5	0.98
United Kingdom	518	2.5	523	2.8	-4.9	0.80
United States	491	3.1	494	3.5	-3.0	0.78
Uruguay	438	2.9	422	3.1	16.6	0.88

Table 13.24 ■ Comparison of problem solving means for all students and problem solving assessed students

	All students		Problem solving assessed students only		Difference	Ratio of error variance
	Mean	SE of mean	Mean	SE of mean	(All - Assessed)	(Assessed/All)
Australia	530	2.0	533	2.2	-3.0	0.84
Austria	506	3.2	508	3.4	-1.8	0.86
Belgium	525	2.2	525	2.3	-0.1	0.94
Brazil	371	4.8	369	5.0	2.2	0.94
Canada	529	1.7	530	2.0	-0.8	0.78
Czech Republic	516	3.4	518	3.6	-1.8	0.91
Denmark	517	2.5	520	2.6	-3.1	0.97
Finland	548	1.9	546	2.2	2.0	0.71
France	519	2.7	517	2.8	1.8	0.90
Germany	513	3.2	514	3.5	-0.9	0.85
Greece	448	4.0	448	4.0	0.1	0.96
Hong Kong	548	4.2	548	4.6	0.2	0.84
Hungary	501	2.9	498	3.0	3.1	0.91
Iceland	505	1.4	502	2.0	2.9	0.49
Indonesia	361	3.3	357	3.5	4.4	0.90
Ireland	498	2.3	496	2.9	2.9	0.63
Italy	469	3.1	471	3.3	-2.0	0.91
Japan	547	4.1	545	4.3	1.8	0.90
Korea	550	3.1	547	3.1	3.4	0.96
Latvia	483	3.9	480	3.8	2.1	1.04
Liechtenstein	529	3.9	533	6.1	-3.9	0.42
Luxembourg	494	1.4	498	1.5	-4.0	0.82
Macao-China	532	2.5	530	4.0	2.4	0.40
Mexico	384	4.3	382	4.6	1.9	0.88
Netherlands	520	3.0	524	3.1	-3.3	0.92
New Zealand	533	2.2	533	2.8	-0.5	0.60
Norway	490	2.6	490	3.0	-0.5	0.76
Poland	487	2.8	486	3.0	0.4	0.87
Portugal	470	3.9	468	4.2	1.4	0.87
Russian Federation	479	4.6	479	4.7	-0.2	0.97
Serbia	420	3.3	421	3.5	-0.9	0.90
Slovak Republic	492	3.4	490	3.5	1.7	0.91
Spain	482	2.7	481	2.9	1.5	0.90
Sweden	509	2.4	512	2.7	-3.6	0.81
Switzerland	521	3.0	526	2.8	-4.4	1.21
Thailand	425	2.7	419	2.8	5.7	0.95
Tunisia	345	2.1	349	2.6	-4.4	0.68
Turkey	408	6.0	412	6.2	-4.3	0.93
United Kingdom	510	2.4	512	2.6	-2.2	0.82
United States	477	3.1	480	3.2	-2.8	0.97
Uruguay	411	3.7	413	3.6	-2.2	1.04

In each case these differences can be explained by characteristics of the students who did not respond to items from the respective domain. In Denmark, students performed surprisingly poorly on booklet 9 when responding to both the science and the reading items. In contrast they performed quite well (relative to other booklets) on problem solving. In addition, it has been noted that the non-responding students (for each domain) have a lower value in the index of economic, social and cultural status (ESCS) than students who did respond to items on each domain. Given the positive correlation between ESCS and achievement, the lower values of ESCS for the students who were not assessed in a domain, and the lower than expected scores on booklet 9, it can be expected that the imputations for the non-assessed students will lead to a reduction in the mean scores in reading and science for Denmark.

In the case of Canada, the mean on science of all students is nine points lower than the mean of the assessed students. This is because Canadian students have not performed well on booklet 9. Interestingly, it appears that the fatigue effect that normally results in PISA booklet differences is less pronounced in Canada than in other countries.

For each of Greece, Hungary and Mexico, a higher than expected performance on the reference booklet has resulted in the mean science scores for all students being higher than the mean science scores for the assessed students.

COMPUTATION OF THE LINK ERROR

Link errors (as discussed in Chapter 9) were obtained by estimating the item parameters for PISA 2000 and PISA 2003 using the international calibration samples. Tables 13.25, 13.26, 13.27 and 13.28 show the item parameter estimates for the items that were common to the two studies for reading, science and the two common mathematics scales (space and shape, and change and relationships) respectively.

The column headed "Difference" in each of these tables shows the amount by which the difference between the estimated item parameters differs from the average difference. The standard deviation of these differences divided by the square root of the number of link items gives the standard errors of the differences under the assumption that the link items are a random sample from some universe of possible link items between 2000 and 2003.

The link standard errors in logits, and on the PISA scale, are given in Table 13.29.

Table 13.25 ■ **Comparison of reading item parameters for PISA 2000 and PISA 2003**

Item name	Difficulty estimate 2003	Centred difficulty estimate 2003	Difficulty estimate 2000	Centred difficulty estimate 2000	Difference	Difference squared
R055Q01	-1.28	-1.28	-1.377	-1.347	-0.072	0.005
R055Q02	0.63	0.63	0.496	0.526	-0.101	0.010
R055Q03	0.27	0.27	0.067	0.097	-0.175	0.031
R055Q05	-0.69	-0.69	-0.877	-0.847	-0.154	0.024
R067Q01	-2.08	-2.08	-1.726	-1.696	0.388	0.151
R067Q04	0.25	0.25	0.516	0.546	0.292	0.085
R067Q05	-0.18	-0.18	0.182	0.212	0.394	0.155
R102Q04A	1.53	1.53	1.206	1.236	-0.290	0.084
R102Q05	0.87	0.87	0.905	0.935	0.067	0.005
R102Q07	-1.42	-1.42	-1.566	-1.536	-0.116	0.013
R104Q01	-1.47	-1.47	-1.235	-1.205	0.268	0.072
R104Q02	1.44	1.44	1.105	1.135	-0.306	0.094
R104Q05	2.17	2.17	1.875	1.905	-0.267	0.071
R111Q01	-0.19	-0.19	-0.053	-0.023	0.164	0.027
R111Q02B	1.54	1.54	1.365	1.395	-0.147	0.022
R111Q06B	0.89	0.89	0.808	0.838	-0.051	0.003
R219Q01T	-0.59	-0.59	-0.550	-0.520	0.069	0.005
R219Q01E	0.10	0.10	0.278	0.308	0.210	0.044
R219Q02	-1.13	-1.13	-0.917	-0.887	0.243	0.059
R220Q01	0.86	0.86	0.785	0.815	-0.041	0.002
R220Q02B	-0.14	-0.14	-0.144	-0.114	0.027	0.001
R220Q04	-0.10	-0.10	0.163	0.193	0.297	0.088
R220Q05	-1.39	-1.39	-1.599	-1.569	-0.184	0.034
R220Q06	-0.34	-0.34	-0.172	-0.142	0.196	0.038
R227Q01	0.40	0.40	0.196	0.226	-0.170	0.029
R227Q02T	0.16	0.16	0.045	0.075	-0.086	0.007
R227Q03	0.46	0.46	0.295	0.325	-0.132	0.017
R227Q06	-0.56	-0.56	-0.916	-0.886	-0.327	0.107

Table 13.26 ■ **Comparison of science item parameters for PISA 2000 and PISA 2003**

Item name	Difficulty estimate 2003	Centred difficulty estimate 2003	Difficulty estimate 2000	Centred difficulty estimate 2000	Difference	Difference squared
S114Q03T	-0.29	-0.30	-0.373	-0.346	0.049	0.002
S114Q04T	0.54	0.54	0.377	0.404	0.133	0.018
S114Q05T	1.48	1.47	1.307	1.334	0.139	0.019
S128Q01	-0.66	-0.67	-0.557	-0.530	-0.138	0.019
S128Q02	0.20	0.20	0.284	0.311	-0.116	0.013
S128Q03T	-0.52	-0.53	-0.527	-0.500	-0.030	0.001
S129Q01	0.42	0.42	0.620	0.647	-0.231	0.053
S129Q02T	1.53	1.53	1.497	1.524	0.004	0.000
S131Q02T	0.26	0.26	0.028	0.055	0.201	0.041
S131Q04T	1.41	1.40	1.438	1.465	-0.063	0.004
S133Q01	-0.60	-0.60	-0.356	-0.329	-0.274	0.075
S133Q03	0.64	0.64	0.313	0.340	0.295	0.087
S133Q04T	0.13	0.13	0.250	0.277	-0.151	0.023
S213Q01T	0.36	0.35	0.419	0.446	-0.094	0.009
S213Q02	-1.46	-1.46	-1.484	-1.457	-0.005	0.000
S252Q01	-0.18	-0.19	0.026	0.053	-0.241	0.058
S252Q02	-0.97	-0.97	-1.123	-1.096	0.124	0.015
S252Q03T	-0.46	-0.47	-0.176	-0.149	-0.323	0.104
S256Q01	-2.21	-2.22	-2.491	-2.464	0.245	0.060
S268Q01	-1.10	-1.11	-1.250	-1.223	0.117	0.014
S268Q02T	0.80	0.79	0.578	0.605	0.188	0.035
S268Q06	-0.17	-0.17	-0.236	-0.209	0.034	0.001
S269Q01	-0.46	-0.46	-0.460	-0.433	-0.030	0.001
S269Q03T	0.56	0.55	0.497	0.524	0.026	0.001
S269Q04T	0.89	0.88	0.712	0.739	0.141	0.020

Table 13.27 ■ **Comparison of space and shape item parameters for PISA 2000 and PISA 2003**

Item name	Difficulty estimate 2003	Centred difficulty estimate 2003	Difficulty estimate 2000	Centred difficulty estimate 2000	Difference	Difference squared
M033Q01	-1.52048	-1.496	-1.38728	-1.410	0.022	0.000506
M034Q01T	0.45924	0.432	0.592436	0.518	0.074	0.005508
M144Q01T	-1.01169	-0.666	-0.87849	-0.580	-0.299	0.089232
M144Q02T	1.08967	1.235	1.222866	1.321	-0.098	0.009674
M144Q03	-1.81466	-1.491	-1.68146	-1.405	-0.277	0.076556
M144Q04T	0.43081	0.641	0.564006	0.727	-0.163	0.02664
M145Q01T	-0.5594	-0.906	-0.4262	-0.820	0.394	0.1549
M266Q01T	1.85779	1.782	1.990986	1.868	0.123	0.015071
M273Q01T	-0.13004	-0.307	0.003156	-0.221	0.224	0.050146

Table 13.28 ■ **Comparison of change and relationships item parameters for PISA 2000 and PISA 2003**

Item name	Difficulty estimate 2003	Centred difficulty estimate 2003	Difficulty estimate 2000	Centred difficulty estimate 2000	Difference	Difference squared
M124Q01	0.53645	0.797	0.478116	0.682	-0.204	0.041691
M124Q03T	1.27627	1.488	1.217936	1.373	-0.155	0.024138
M150Q01	-0.68604	-0.913	-0.74437	-1.028	0.283	0.080274
M150Q02T	-1.12923	-0.979	-1.18756	-1.094	-0.094	0.00881
M150Q03T	0.00896	0.322	-0.04937	0.207	-0.257	0.065882
M155Q01	-0.74461	-0.891	-0.80294	-1.006	0.203	0.04111
M155Q02T	-0.64269	-0.480	-0.70102	-0.595	-0.106	0.011305
M155Q03T	1.71785	1.616	1.659516	1.501	0.158	0.025032
M155Q04T	-0.23021	-0.391	-0.28854	-0.506	0.217	0.047157
M192Q01T	0.47659	0.578	0.418256	0.463	-0.045	0.002029

Table 13.29 ■ **Standard errors for the PISA 2000 to PISA 2003 links**

Scale	Standard error on logits	Standard error on PISA scale
Reading	0.041	3.744
Science	0.033	2.959
Space and shape	0.077	6.008
Change and relationships	0.062	4.84

TRANSFORMING THE PLAUSIBLE VALUES TO PISA SCALES

As described in Chapter 9 the PISA 2003 reporting scales for reading and science are the same as those used in PISA 2000. For mathematics and problem solving new scales were prepared for PISA 2003. The transformations for mapping the PISA 2003 logits to the PISA reporting scales are given below for each domain.

Reading

After computing the plausible values, on the logit metric, and following the procedures described in Chapter 9, it was noted that there were substantial differences between the optimal linking transformations for male and female students. The resulting transformations were as follows:

For male students:

$$P_{2000} = \left\{ \frac{\left[\left(0.8823 I_{2003} + 0.0204 \right) - 0.5076 \right]}{1.1002} \times 100 + 500 \right\} \qquad (13.1)$$

For female students:

$$P_{2000} = \left\{ \frac{\left[\left(0.8739 I_{2003} + 0.0970 \right) - 0.5076 \right]}{1.1002} \times 100 + 500 \right\} \qquad (13.2)$$

For students with missing gender code:

$$P_{2000} = \left\{ \frac{\left[\left(0.8830 l_{2003} + 0.0552 \right) - 0.5076 \right]}{1.1002} \times 100 + 500 \right\} \tag{13.3}$$

The coefficients 0.5076, 1.1002, 100 and 500 are required to transform the PISA 2000 logits to the PISA 2000 scale. The scale factors of 0.8823, 0,8739 and 0.8830 and shifts of 0.0204, 0.0970 and 0.0552 transform the PISA 2003 logit scale to the PISA 2000 logit scale for males, females, and missing gender code students respectively.

Science

For science the transformation is given by:

$$P_{2000} = \left\{ \frac{\left[\left(1.0063 l_{2003} + 0.0155 \right) + 0.0933 \right]}{1.1086} \times 100 + 500 \right\} \tag{13.4}$$

The multiplication by 1.0063 and addition of 0.0155 transforms the 2003 logits to the 2000 logit scale, and then the 2000 logit is transformed to the PISA 2000 scale.

Problem solving

For problem solving the transformations are simpler because they do not involve the transformation of the PISA 2003 logits to the PISA 2000 scale.

$$P_{2003} = \left\{ \frac{\left[l_{2003} + 0.0973 \right]}{1.1751} \times 100 + 500 \right\} \tag{13.5}$$

Mathematics

Similarly for mathematics the transformations are simpler because they do not involve the transformation of the PISA 2003 logits to the PISA 2000 scale.

$$P_{2003} = \left\{ \frac{\left[l_{2003} + 0.1344 \right]}{1.2838} \times 100 + 500 \right\} \tag{13.6}$$

Space and shape

$$P_{2003} = \left\{ \frac{\left[\left(0.996 l_{2000} + 0.008 \right) + 0.1342 \right]}{1.2837} \times 100 + 500 \right\} \tag{13.7}$$

Change and relationships

$$P_{2003} = \left\{ \frac{\left[\left(0.985 l_{2000} + 0.059 \right) + 0.1342 \right]}{1.2837} \times 100 + 500 \right\} \qquad (13.8)$$

Notes

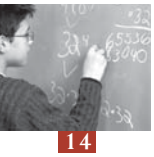
1 Note that both Luxembourg and the United Kingdom have been excluded from these calculations.

2 For the definition of "not reached" see Chapter 18.

3 Note that because the design was balanced the inclusion of the booklet term in the item response model did not have an appreciable effect on the item parameter estimates.

14

Outcomes of Coder Reliability Studies



This chapter reports the result of the various coder reliability studies that were implemented. The methodologies for these studies are described in Chapter 10.

WITHIN-COUNTRY RELIABILITY STUDIES

Variance components analysis

Tables 14.1 to 14.4 show the results of the variance components analysis for the multiple-marked items in mathematics, science, reading and problem solving, respectively. The variance components are each expressed as a percentage of their sum.

The tables show that those variance components associated with markers are remarkably small relative to the other components. This means that there are no significant systematic within-country marker effects.

As discussed in Chapter 10, analyses of the type reported here can result in negative variance estimates. If the amount by which the component is negative is small, then this is a sign that the variance component is negligible (near zero). If the component is large and negative, then it is a sign that the analysis method is inappropriate for the data. Some sub-regions within countries were considered as countries for these analyses. Brazil did not submit their multiple-marked data on time and did not follow the specified coding design. Therefore Brazil has been omitted altogether.

Generalisability co-efficients

The generalisability co-efficients are computed from the variance components using:

$$\rho_3\left(Y_{v..}, Y'_{v..}\right) = \frac{\sigma_A^2 + \frac{\sigma^2_{AB+E^*}}{I}}{\sigma_A^2 + \frac{\sigma^2_{AB+E^*}}{I} + \frac{\sigma^2_{ac}}{R} + \frac{\sigma^2_{abc+e^*}}{I \times R}} \qquad (14.1)$$

and

$$\rho_3\left(Y_{v..}, Y'_{v..}\right) = \frac{\sigma_A^2 + \frac{\sigma^2_{AB}}{I}}{\sigma_A^2 + \frac{\sigma^2_{AB} + \sigma^2_{\varepsilon^*}}{I} + \frac{\sigma^2_{ac}}{R} + \frac{\sigma^2_{abc+e^*}}{I \times R}} \qquad (14.2)$$

They provide an index of reliability for the multiple coding in each country. *I* denotes the number of items and *M* the number of coders. By using different values for *I* and *M*, one obtains a generalisation of the Spearman-Brown formula for test-lengthening. In Tables 14.5 to 14.8, the formula is evaluated for the three combinations of $I = \{8, 16, 24\}$ and $M = 1$, using the variance component estimates from the corresponding tables presented above. For the countries marked with an asterisk ('*') in the above tables, no values are displayed, because they fall outside the acceptable (0,1) range.

Table 14.1 ■ **Variance components for mathematics coding**

	Student	Item	Coder	Student-item interaction	Student-coder interaction	Item-coder interaction	Measurement error
Australia	17.08	30.41	0.01	46.72	0.02	0.07	5.69
Austria	21.37	24.04	0.01	50.06	0.01	0.05	4.45
Belgium (Flemish)	18.24	32.06	0.03	41.93	0.10	0.31	7.31
Belgium (French)	20.73	26.19	0.00	48.43	0.01	0.06	4.57
Canada (English)	15.09	32.20	0.07	43.87	-8.56	0.37	16.96
Canada (French)	14.09	28.07	0.03	49.03	-0.14	0.25	8.68
Czech Republic	22.36	24.66	0.03	48.02	0.00	0.11	4.82
Denmark	12.71	26.91	0.06	50.32	-0.18	0.43	9.75
England	17.74	32.36	0.01	46.27	0.04	0.04	3.55
Finland	15.26	28.30	0.00	54.11	0.01	0.02	2.30
France	15.33	31.07	-0.01	45.66	-0.72	0.23	8.43
Germany	20.46	26.25	0.02	47.30	0.10	0.21	5.67
Greece	18.44	24.19	0.05	49.64	0.05	0.21	7.41
Hong Kong-China	16.99	30.25	0.00	46.86	0.00	0.20	5.69
Hungary	18.33	27.13	-0.01	48.20	0.09	0.23	6.02
Iceland	16.20	29.86	-0.01	49.24	0.02	0.10	4.58
Indonesia	15.48	13.23	0.00	69.67		0.00	1.60
Ireland	14.68	32.25	-0.01	44.35	0.01	0.20	8.51
Italy	16.81	29.14	0.00	49.72	0.03	0.07	4.23
Italy (German)	14.46	31.35	0.02	49.44	-0.09	0.07	4.75
Japan	19.94	25.63	0.00	53.59	-1.14	-0.02	2.00
Korea	17.79	30.63	0.01	48.46	-0.01	0.05	3.07
Latvia	17.45	28.12	-0.06	44.38	0.50	1.04	8.58
Luxembourg	18.48	25.75	0.01	51.14	0.09	0.12	4.42
Mexico	13.62	21.88	0.04	54.74	0.18	0.25	9.30
Netherlands	16.16	35.52	0.02	41.11	0.01	0.12	7.06
New Zealand	18.07	30.41	0.05	43.26	0.01	0.15	8.05
Norway	14.37	33.85	0.00	45.57	0.08	0.17	5.96
Poland	21.36	24.76	0.00	51.25	0.02	0.02	2.59
Portugal	15.18	34.04	0.00	49.71	0.03	0.00	1.04
Russian Federation	17.77	30.45	0.00	45.65	0.08	0.13	5.91
Scotland	18.53	29.86	0.00	48.42	0.04	0.02	3.13
Serbia	18.42	20.62	0.01	55.63	0.01	0.07	5.25
Slovak Republic	19.31	24.65	0.00	50.53	0.10	0.06	5.35
Spain (Basque)	16.23	33.01	0.00	49.93	0.01	0.00	0.83
Spain (Catalan)	15.33	31.79	0.11	45.43	0.04	0.38	6.92
Spain (Leon)	15.44	33.56	0.22	43.12	0.23	0.67	6.74
Spain (Other)	16.82	28.06	0.06	49.19	0.21	0.25	5.41
Sweden	20.78	25.88	0.00	46.47	0.05	0.33	6.49
Switzerland (French)	16.07	27.22	0.01	49.79	0.11	0.14	6.67
Switzerland (German)	19.90	24.11	0.04	47.81	0.00	0.23	7.91
Thailand	16.44	29.01	0.01	51.81	0.07	0.03	2.64
Tunisia	19.40	14.28	0.21	49.13	-13.31	0.69	29.59
Turkey	18.08	24.05	0.00	57.48	-0.01	0.00	0.39
United States	17.22	32.66	0.01	41.06	-5.69	0.16	14.59
Uruguay	16.87	23.29	0.00	53.69	-0.01	0.11	6.03

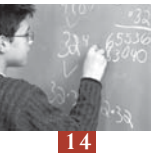


Table 14.2 ■ **Variance components for science coding**

	Student	Item	Coder	Student-item interaction	Student-coder interaction	Item-coder interaction	Measure-ment error
Australia	14.82	19.07	0.01	52.77	-0.04	0.26	13.11
Austria	19.71	12.06	0.11	54.20	0.07	0.31	13.55
Belgium (Flemish)	27.19	8.34	0.13	49.23	-0.03	0.34	14.80
Belgium (French)	26.83	11.53	0.04	51.17	0.19	0.08	10.16
Canada (English)	25.33	2.44	0.07	66.37	0.09	0.02	5.68
Canada (French)	25.70	2.39	-0.01	67.56	0.10	0.09	4.17
Czech Republic	22.14	8.47	0.03	61.08	0.03	0.52	7.74
Denmark	26.92	4.71	0.00	66.32	-0.06	0.02	2.08
England	21.16	17.11	-0.01	56.06	0.08	0.02	5.58
Finland	17.29	16.21	0.00	61.65	-0.05	0.01	4.88
France	22.82	12.16	-0.05	46.14	0.04	0.90	17.98
Germany	23.72	12.23	-0.03	49.10	0.14	0.43	14.41
Greece	21.59	13.09	0.01	60.72	0.00	0.06	4.53
Hong Kong-China	16.62	18.18	0.09	48.59	0.18	0.34	15.99
Hungary	22.19	11.29	0.05	56.16	0.17	0.04	10.10
Iceland	17.17	19.31	0.04	53.42	0.02	0.16	9.89
Indonesia	18.21	3.97	0.01	75.26	0.26	0.01	2.29
Ireland	18.76	11.72	-0.04	49.36	0.10	0.85	19.23
Italy	18.38	12.95	0.03	62.17	0.00	0.08	6.39
Italy (German)	13.88	18.08	0.28	57.22	0.06	0.27	10.21
Japan	25.12	18.94	0.00	55.00	0.00	0.00	0.94
Korea	22.84	11.67	0.00	56.42	0.00	0.10	8.97
Latvia	14.35	14.47	0.31	54.96	0.22	0.68	15.01
Luxembourg	22.98	12.02	0.00	57.26	0.02	0.04	7.69
Mexico	16.50	4.94	0.66	50.78	0.14	1.11	25.88
Netherlands	19.23	13.79	0.07	52.31	-0.04	0.65	14.00
New Zealand	18.68	13.84	0.22	44.35	-0.06	0.49	22.47
Norway	20.41	15.31	0.04	47.99	-0.01	0.56	15.70
Poland	23.77	8.45	0.02	62.00	0.01	0.03	5.72
Portugal	16.81	14.17	0.00	68.22	0.00	0.01	0.80
Russian Federation	17.71	14.64	0.00	66.07	0.02	0.01	1.55
Scotland	24.68	13.41	0.01	55.32	-0.02	0.05	6.55
Serbia	19.43	8.44	0.08	52.6	-1.48	0.76	20.16
Slovak Republic	22.26	10.66	0.00	54.36	-0.04	0.06	12.69
Spain (Basque)	24.46	8.62	0.28	51.78	-0.02	0.81	14.07
Spain (Catalan)	24.46	8.62	0.28	51.78	-0.02	0.81	14.07
Spain (Leon)	15.63	10.09	0.51	56.65	0.26	1.32	15.56
Spain (Other)	19.55	10.44	0.16	50.23	0.07	1.27	18.29
Sweden	27.42	9.65	0.06	51.38	0.00	0.22	11.27
Switzerland (French)	21.82	7.33	0.19	53.68	-0.09	1.02	16.05
Switzerland (German)	23.26	10.59	0.17	50.71	0.05	0.30	14.91
Thailand	87.13	1.83	0.00	9.21	0.00	0.03	1.80
Tunisia	18.61	7.23	0.66	41.38	-8.32	2.4	38.03
Turkey	18.65	8.84	0.00	69.76	-0.03	0.00	2.79
United States	18.49	15.24	-0.01	57.25	0.09	0.09	8.85
Uruguay	18.99	7.96	0.22	54.33	0.44	0.23	17.84

© OECD 2005

Table 14.3 ■ **Variance components for reading coding**

	Student	Item	Coder	Student-item interaction	Student-coder interaction	Item-coder interaction	Measure-ment error
Australia	19.86	23.63	0.05	44.85	0.12	0.14	11.33
Austria	15.06	12.24	0.06	55.05	-0.04	0.42	17.22
Belgium (Flemish)	16.76	22.33	-0.01	45.59	0.12	0.01	15.19
Belgium (French)	19.90	23.63	-0.04	43.04	0.12	0.58	12.76
Canada (English)	18.11	2.61	0.00	73.65	-0.02	0.10	5.54
Canada (French)	7.90	4.02	0.01	76.51	0.05	0.02	11.49
Czech Republic	12.50	28.93	0.07	46.76	-0.04	0.37	11.40
Denmark	18.12	3.32	0.02	74.43	0.02	0.14	3.96
England	22.44	26.05	0.02	42.81	0.07	0.07	8.55
Finland	16.03	21.13	0.05	52.03	-0.06	0.10	10.73
France	22.46	19.27	0.01	41.03	-0.08	0.51	16.8
Germany	17.98	12.17	0.08	55.12	0.04	0.46	14.16
Greece	14.73	28.36	-0.08	43.41	0.15	0.99	12.45
Hong Kong-China	15.65	28.85	0.01	44.17	-0.07	0.13	11.27
Hungary	14.89	18.14	0.06	51.83	0.08	0.22	14.77
Iceland	16.34	16.93	0.07	54.04	-0.01	0.46	12.18
Indonesia	8.82	20.82	0.00	69.08	-0.02	0.00	1.30
Ireland	15.29	26.37	-0.04	42.49	0.09	0.71	15.09
Italy	15.65	26.98	0.00	46.85	-0.07	0.38	10.20
Italy (German)	15.63	20.33	0.20	47.36	-0.63	0.28	16.83
Japan	21.23	23.43	-0.09	50.12	-9.49	-0.15	14.95
Korea	14.35	31.49	-0.01	44.84	0.06	0.23	9.04
Latvia	13.58	27.80	0.08	47.85	0.18	0.21	10.30
Luxembourg	18.80	16.59	0.02	55.74	0.07	0.00	8.78
Mexico	10.45	30.82	0.27	39.55	0.22	0.98	17.72
Netherlands	17.11	18.11	-0.03	49.64	0.11	0.52	14.53
New Zealand	21.63	21.50	0.00	41.40	0.04	0.23	15.21
Norway	23.24	15.42	0.05	48.74	0.03	0.21	12.31
Poland	23.45	21.17	0.00	51.78	0.00	0.02	3.57
Portugal	17.06	31.27	0.00	50.77	0.00	0.00	0.90
Russian Federation	14.38	17.94	0.00	64.92	0.01	0.00	2.76
Scotland	13.57	25.80	0.01	51.50	-0.03	0.03	9.12
Serbia	19.15	10.53	0.00	60.85	0.04	0.20	9.23
Slovak Republic	11.15	25.98	0.24	41.35	0.05	0.82	20.41
Spain (Basque)	13.19	31.77	-0.04	49.10	0.00	0.10	5.89
Spain (Catalan)	13.03	32.49	-0.02	43.02	-0.01	0.06	11.43
Spain (Leon)	13.84	29.88	0.11	44.47	-0.02	0.25	11.48
Spain (Other)	14.76	32.25	0.06	41.69	0.16	0.31	10.78
Sweden	22.92	22.26	0.01	41.61	0.15	0.13	12.92
Switzerland (French)	17.71	19.45	1.35	46.98	-9.17	-0.68	24.37
Switzerland (German)	19.44	13.74	0.10	52.81	0.16	0.15	13.61
Thailand	77.10	5.95	0.00	14.85	-0.01	0.03	2.08
Tunisia	17.56	15.26	0.01	57.06	0.14	0.18	9.80
Turkey	16.29	23.27	0.00	58.68	0.07	0.00	1.69
United States	16.56	26.13	0.01	50.62	-0.03	0.06	6.66
Uruguay	17.49	22.02	0.14	44.60	0.10	1.17	14.46

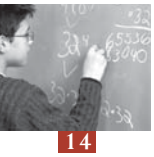


Table 14.4 ■ **Variance components for problem-solving coding**

	Student	Item	Coder	Student-item interaction	Student-coder interaction	Item-coder interaction	Measure-ment error
Australia	31.43	2.50	0.01	62.80	0.01	0.02	3.23
Austria	28.99	3.54	0.00	64.65	-0.04	0.02	2.85
Belgium (Flemish)	28.14	4.36	0.14	58.07	0.13	0.49	8.67
Belgium (French)	29.67	0.47	-0.01	67.06	0.02	0.06	2.72
Canada (English)	8.82	3.64	0.00	82.00	-0.02	0.01	5.55
Canada (French)	28.02	2.22	0.00	64.99	0.15	0.03	4.60
Czech Republic	28.96	1.98	0.02	65.46	-0.07	0.01	3.64
Denmark	18.32	7.01	0.01	72.45	0.01	0.00	2.19
England	34.64	1.83	-0.01	59.37	0.04	0.06	4.07
Finland	30.34	6.16	0.00	61.04	-0.01	0.01	2.46
France	24.67	3.91	-0.04	62.88	-1.33	-0.02	9.92
Germany	28.33	3.40	0.00	64.04	0.09	0.00	4.14
Greece	29.74	6.42	0.01	60.26	0.15	0.01	3.42
Hong Kong-China	32.71	4.79	0.02	57.67	0.04	0.12	4.65
Hungary	36.21	3.60	0.05	53.47	0.00	0.28	6.39
Iceland	23.54	5.05	0.01	66.24	-0.04	0.03	5.17
Indonesia	7.35	4.93	0.00	86.81	0.01	0.00	0.90
Ireland	22.89	7.16	-0.03	61.49	-0.01	0.25	8.24
Italy	26.58	7.01	0.01	63.67	-0.12	0.02	2.83
Italy (German)	22.68	6.01	-0.01	65.61	0.00	0.05	5.67
Japan	32.95	5.91	0.00	60.61	0.00	0.00	0.53
Korea	27.62	5.22	0.00	62.63	0.00	0.08	4.45
Latvia	22.13	10.23	0	61.35	0.04	0.15	6.09
Luxembourg	20.84	4.03	0.00	70.20	0.09	0.03	4.81
Mexico	17.05	8.58	0.25	48.19	0.28	2.94	22.71
Netherlands	21.77	4.03	-0.03	64.51	-0.15	0.24	9.63
New Zealand	31.52	3.35	0.01	57.45	0.11	0.14	7.42
Norway	28.85	4.15	0.01	61.32	0.07	0.05	5.55
Poland	24.51	6.66	0.00	64.67	-0.08	0.02	4.22
Portugal	27.17	7.32	0.00	64.73	0.00	0.02	0.76
Russian Federation	29.67	6.84	0.00	60.57	-0.01	0.00	2.93
Scotland	30.78	1.19	0.00	64.96	-0.07	0.00	3.14
Serbia	21.93	3.25	0.08	69.4	-3.91	-0.18	9.45
Slovak Republic	22.94	6.38	0.03	58.66	0.13	0.31	11.55
Spain (Basque)	20.19	12.17	0.00	66.78	0.01	0.00	0.85
Spain (Catalan)	25.60	8.77	0.01	56.21	0.18	0.46	8.77
Spain (Leon)	10.64	4.19	0.00	83.58	0.00	0.03	1.57
Spain (Other)	20.65	8.15	-0.04	58.27	0.22	0.81	11.94
Sweden	26.63	3.82	0.07	61.87	0.13	0.15	7.33
Switzerland (French)	31.99	3.13	-0.05	58.95	0.05	0.52	5.42
Switzerland (German)	19.42	4.41	-0.01	70.78	-0.04	0.15	5.29
Thailand	92.81	0.96	0.00	5.56	0.01	0.01	0.66
Tunisia	12.93	4.35	0.24	45.2	-15.28	0.84	51.71
Turkey	30.65	4.76	0.00	63.82	-0.01	0.00	0.78
United States	32.43	3.45	0.01	61.44	-3.00	-0.13	5.80
Uruguay	26.52	3.82	-0.01	61.83	-0.14	0.09	7.89

© OECD 2005

Table 14.5 ■ **Estimates for mathematics inter-coder reliability**

	I = 8, M = 1		I = 16, M = 1		I = 24, M = 1	
	ρ_3	ρ_4	ρ_3	ρ_4	ρ_3	ρ_4
Australia	0.969	0.722	0.982	0.838	0.987	0.886
Austria	0.980	0.758	0.988	0.862	0.992	0.904
Belgium (Flemish)	0.959	0.745	0.974	0.852	0.980	0.894
Belgium (French)	0.979	0.758	0.988	0.862	0.991	0.903
Canada (French)	0.955	0.666	0.977	0.803	0.986	0.862
Czech Republic	0.979	0.772	0.988	0.871	0.992	0.910
Denmark	0.948	0.634	0.974	0.781	0.985	0.845
England	0.980	0.739	0.987	0.849	0.991	0.893
Finland	0.987	0.684	0.992	0.812	0.994	0.866
France	0.984	0.717	1.011	0.852	1.022	0.909
Germany	0.970	0.753	0.981	0.857	0.985	0.899
Greece	0.962	0.720	0.977	0.836	0.983	0.884
Hong Kong-China	0.970	0.721	0.982	0.838	0.988	0.886
Hungary	0.967	0.727	0.979	0.840	0.984	0.886
Iceland	0.974	0.706	0.984	0.827	0.989	0.877
Indonesia	0.991	0.634	0.994	0.776	0.995	0.838
Ireland	0.950	0.689	0.970	0.816	0.978	0.869
Italy	0.976	0.713	0.985	0.832	0.989	0.881
Italy (German)	0.976	0.684	0.988	0.814	0.994	0.870
Korea	0.985	0.734	0.991	0.847	0.994	0.893
Latvia	0.936	0.710	0.951	0.821	0.957	0.866
Luxembourg	0.975	0.724	0.983	0.838	0.987	0.885
Mexico	0.938	0.625	0.957	0.765	0.966	0.827
Netherlands	0.960	0.728	0.976	0.843	0.983	0.889
New Zealand	0.959	0.738	0.976	0.849	0.983	0.894
Norway	0.961	0.688	0.974	0.813	0.980	0.866
Poland	0.988	0.760	0.993	0.863	0.995	0.904
Portugal	0.993	0.704	0.995	0.826	0.996	0.876
Russian Federation	0.966	0.731	0.979	0.843	0.984	0.889
Scotland	0.983	0.741	0.989	0.850	0.992	0.894
Serbia	0.974	0.707	0.985	0.828	0.989	0.879
Slovak Republic	0.971	0.732	0.981	0.843	0.985	0.888
Spain (Basque)	0.995	0.719	0.997	0.836	0.998	0.884
Spain (Catalan)	0.959	0.700	0.975	0.822	0.981	0.873
Spain (Leon)	0.951	0.705	0.965	0.822	0.971	0.870
Spain (Other)	0.963	0.705	0.973	0.823	0.977	0.871
Sweden	0.969	0.757	0.981	0.861	0.986	0.902
Switzerland (French)	0.959	0.692	0.973	0.815	0.979	0.867
Switzerland (German)	0.963	0.741	0.979	0.851	0.985	0.896
Thailand	0.983	0.705	0.988	0.826	0.990	0.875
Turkey	0.998	0.715	0.999	0.834	1.000	0.883
Uruguay	0.969	0.694	0.982	0.819	0.988	0.872

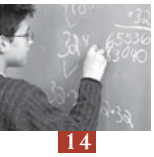


Table 14.6 ■ **Estimates for science inter-coder reliability**

	I = 8, M = 1		I = 16, M = 1		I = 24, M = 1	
	ρ_3	ρ_4	ρ_3	ρ_4	ρ_3	ρ_4
Australia	0.930	0.644	0.959	0.784	0.971	0.846
Austria	0.938	0.698	0.962	0.821	0.972	0.872
Belgium (Flemish)	0.948	0.773	0.971	0.873	0.980	0.912
Belgium (French)	0.958	0.774	0.973	0.870	0.979	0.907
Canada (English)	0.977	0.736	0.985	0.846	0.988	0.891
Canada (French)	0.982	0.739	0.988	0.849	0.990	0.893
Czech Republic	0.968	0.720	0.981	0.836	0.986	0.884
Denmark	0.994	0.760	0.998	0.865	0.999	0.906
England	0.973	0.731	0.983	0.843	0.987	0.889
Finland	0.978	0.676	0.988	0.808	0.992	0.864
France	0.926	0.739	0.957	0.849	0.969	0.894
Germany	0.939	0.746	0.963	0.852	0.972	0.895
Greece	0.981	0.726	0.989	0.841	0.992	0.888
Hong Kong-China	0.912	0.668	0.943	0.798	0.956	0.853
Hungary	0.953	0.724	0.970	0.837	0.976	0.883
Iceland	0.950	0.684	0.970	0.812	0.978	0.866
Indonesia	0.981	0.647	0.983	0.781	0.984	0.839
Ireland	0.909	0.684	0.944	0.810	0.958	0.864
Italy	0.970	0.682	0.982	0.811	0.987	0.865
Italy (German)	0.940	0.621	0.962	0.765	0.971	0.829
Japan	0.996	0.782	0.998	0.878	0.999	0.915
Korea	0.964	0.736	0.979	0.848	0.985	0.893
Latvia	0.910	0.616	0.939	0.758	0.952	0.821
Luxembourg	0.969	0.739	0.982	0.849	0.987	0.894
Mexico	0.871	0.629	0.918	0.770	0.939	0.832
The Netherlands	0.938	0.700	0.964	0.824	0.975	0.876
New Zealand	0.898	0.693	0.941	0.819	0.959	0.873
Norway	0.931	0.720	0.960	0.837	0.972	0.885
Poland	0.977	0.737	0.987	0.848	0.990	0.893
Portugal	0.996	0.661	0.998	0.796	0.998	0.854
Russian Federation	0.992	0.676	0.995	0.807	0.996	0.862
Scotland	0.975	0.762	0.986	0.865	0.991	0.906
Slovak Republic	0.949	0.727	0.971	0.843	0.980	0.890
Spain (Basque)	0.947	0.749	0.970	0.857	0.979	0.900
Spain (Catalan)	0.947	0.749	0.970	0.857	0.979	0.900
Spain (Leon)	0.912	0.627	0.940	0.766	0.952	0.827
Spain (Other)	0.917	0.694	0.949	0.818	0.963	0.870
Sweden	0.960	0.778	0.978	0.875	0.985	0.913
Switzerland (French)	0.937	0.717	0.965	0.837	0.977	0.886
Switzerland (German)	0.939	0.738	0.964	0.849	0.974	0.893
Thailand	0.997	0.984	0.999	0.992	0.999	0.995
Turkey	0.988	0.673	0.994	0.805	0.996	0.862
United States	0.956	0.689	0.972	0.814	0.979	0.867
Uruguay	0.906	0.668	0.935	0.793	0.947	0.846

Table 14.7 ■ **Estimates for reading inter-coder reliability**

	I = 8, M = 1		I = 16, M = 1		I = 24, M = 1	
	ρ_3	ρ_4	ρ_3	ρ_4	ρ_3	ρ_4
Australia	0.943	0.735	0.965	0.845	0.973	0.890
Austria	0.912	0.626	0.947	0.771	0.962	0.835
Belgium (Flemish)	0.917	0.685	0.948	0.810	0.961	0.863
Belgium (French)	0.936	0.737	0.961	0.846	0.971	0.890
Canada (English)	0.976	0.647	0.986	0.786	0.990	0.847
Canada (French)	0.922	0.417	0.943	0.587	0.954	0.680
Czech Republic	0.930	0.634	0.958	0.777	0.971	0.840
Denmark	0.981	0.649	0.988	0.786	0.991	0.846
England	0.961	0.776	0.977	0.873	0.983	0.910
Finland	0.946	0.673	0.969	0.806	0.979	0.862
France	0.932	0.759	0.963	0.864	0.975	0.906
Germany	0.932	0.674	0.958	0.804	0.970	0.860
Greece	0.922	0.674	0.949	0.802	0.961	0.856
Hong Kong-China	0.940	0.695	0.967	0.822	0.978	0.875
Hungary	0.917	0.639	0.947	0.778	0.961	0.839
Iceland	0.939	0.664	0.963	0.798	0.974	0.856
Indonesia	0.992	0.501	0.996	0.668	0.997	0.752
Ireland	0.913	0.677	0.946	0.806	0.960	0.860
Italy	0.947	0.689	0.970	0.817	0.980	0.872
Italy (German)	0.936	0.679	0.978	0.822	0.996	0.884
Korea	0.944	0.679	0.965	0.807	0.974	0.862
Latvia	0.930	0.646	0.953	0.781	0.962	0.839
Luxembourg	0.957	0.698	0.973	0.821	0.980	0.872
Mexico	0.864	0.586	0.907	0.734	0.927	0.801
Netherlands	0.924	0.678	0.952	0.806	0.964	0.860
New Zealand	0.933	0.753	0.961	0.858	0.972	0.900
Norway	0.949	0.752	0.971	0.858	0.979	0.900
Poland	0.985	0.772	0.992	0.871	0.994	0.910
Portugal	0.995	0.725	0.997	0.841	0.998	0.888
Russian Federation	0.984	0.629	0.990	0.772	0.993	0.835
Scotland	0.947	0.643	0.969	0.783	0.978	0.845
Serbia	0.957	0.685	0.974	0.812	0.981	0.866
Slovak Republic	0.863	0.590	0.912	0.741	0.935	0.810
Spain (Basque)	0.963	0.657	0.978	0.793	0.984	0.852
Spain (Catalan)	0.928	0.657	0.957	0.793	0.970	0.852
Spain (Leon)	0.932	0.665	0.960	0.799	0.972	0.857
Spain (Other)	0.930	0.687	0.954	0.811	0.965	0.863
Sweden	0.941	0.767	0.964	0.865	0.973	0.904
Switzerland (German)	0.933	0.697	0.957	0.819	0.968	0.869
Thailand	0.997	0.973	0.998	0.987	0.999	0.991
Tunisia	0.948	0.674	0.966	0.803	0.973	0.857
Turkey	0.988	0.681	0.991	0.809	0.993	0.863
United States	0.966	0.699	0.981	0.824	0.987	0.875
Uruguay	0.923	0.700	0.953	0.822	0.965	0.872

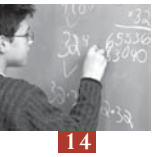


Table 14.8 ■ **Estimates for problem solving inter-coder reliability**

	I = 8, M = 1		I = 16, M = 1		I = 24, M = 1	
	ρ_3	ρ_4	ρ_3	ρ_4	ρ_3	ρ_4
Australia	0.990	0.792	0.994	0.884	0.996	0.919
Austria	0.992	0.775	0.996	0.874	0.998	0.913
Belgium (Flemish)	0.967	0.769	0.979	0.867	0.984	0.906
Belgium (French)	0.991	0.772	0.994	0.871	0.996	0.910
Canada (English)	0.966	0.447	0.977	0.618	0.983	0.708
Canada (French)	0.980	0.760	0.987	0.862	0.989	0.902
Czech Republic	0.990	0.772	0.995	0.872	0.997	0.912
Denmark	0.990	0.662	0.994	0.797	0.995	0.854
England	0.987	0.813	0.992	0.896	0.994	0.928
Finland	0.992	0.793	0.996	0.885	0.997	0.920
Germany	0.984	0.767	0.989	0.867	0.992	0.906
Greece	0.985	0.786	0.989	0.878	0.991	0.914
Hong Kong-China	0.985	0.807	0.991	0.893	0.993	0.925
Hungary	0.982	0.829	0.990	0.906	0.993	0.936
Iceland	0.981	0.726	0.990	0.842	0.994	0.889
Indonesia	0.993	0.401	0.995	0.572	0.996	0.667
Ireland	0.968	0.724	0.981	0.840	0.987	0.888
Italy	0.993	0.765	0.998	0.868	1.000	0.909
Italy (German)	0.978	0.718	0.987	0.836	0.991	0.884
Japan	0.998	0.812	0.999	0.896	0.999	0.928
Korea	0.985	0.767	0.991	0.868	0.994	0.908
Latvia	0.974	0.723	0.984	0.839	0.988	0.886
Luxembourg	0.977	0.688	0.985	0.813	0.988	0.866
Mexico	0.881	0.651	0.922	0.784	0.940	0.841
Netherlands	0.966	0.705	0.983	0.829	0.990	0.881
New Zealand	0.974	0.793	0.984	0.883	0.988	0.918
Norway	0.980	0.774	0.987	0.872	0.991	0.910
Poland	0.986	0.742	0.994	0.853	0.996	0.898
Portugal	0.997	0.768	0.998	0.869	0.999	0.909
Russian Federation	0.991	0.789	0.995	0.882	0.997	0.918
Scotland	0.992	0.785	0.996	0.880	0.998	0.918
Slovak Republic	0.951	0.720	0.969	0.836	0.977	0.883
Spain (Basque)	0.996	0.705	0.998	0.827	0.998	0.877
Spain (Catalan)	0.962	0.755	0.976	0.858	0.981	0.899
Spain (Leon)	0.991	0.500	0.994	0.667	0.995	0.750
Spain (Other)	0.942	0.697	0.962	0.818	0.970	0.868
Sweden	0.970	0.752	0.981	0.857	0.985	0.898
Switzerland (French)	0.982	0.798	0.989	0.887	0.992	0.921
Switzerland (German)	0.979	0.672	0.988	0.805	0.992	0.861
Thailand	0.999	0.992	0.999	0.996	1.000	0.997
Turkey	0.998	0.792	0.999	0.884	0.999	0.920
Uruguay	0.976	0.756	0.988	0.863	0.993	0.905

© OECD 2005

OUTCOME OF THE INTER-COUNTRY CODER RELIABILITY (ICR) STUDY

Some 7 200 booklets were submitted for the ICR study, in which 71 941 student answers were recoded by the verifiers.[1] In about 80 per cent of these cases, both the verifier and all four national coders agreed on an identical code, and in another 7 per cent of the cases the verifier agreed with the majority – that is, three out of four – of the national coders.

Of the remaining cases, about 3 per cent had national codes considered too inconsistent to allow comparison with the verifier's code, and 10 per cent were flagged and submitted to consortium staff for adjudication. In approximately 5 per cent of the adjudicated cases, the adjudicators found that the codes given by the national coders were correct, while 3.1 per cent of the codes were found to be too lenient and 1.8 per cent too harsh.

Hence, as summarised in Table 14.9 a very high proportion of the cases (92 per cent) showed consistency with the scoring instructions described in the coding guides.

Table 14.9 ■ **Summary of the 2003 inter-country coder (ICR) reliability study**

	Mathematics	Science	Reading	Overall
Number of student responses	24 571	23 570	24 600	73 741
% agreement	95.1	90.1	90.5	91.9
% too inconsistent codes	2.1	3.6	3.6	3.1
% too harsh codes	0.9	2.5	2.1	1.8
% too lenient codes	1.8	3.8	3.8	3.2

While relatively infrequent, the inconsistent or biased codes were not uniformly distributed across items or countries, suggesting that scoring instructions may have been insufficiently stringent for some of the items, and that some of the national scoring teams may have been less accurate than others.

Results by item

Table 14.10 presents the detail of the results by item. Items that had less than 90 per cent of cases in the agreement category, and/or had more than 5 per cent of cases assessed as either too harsh or too lenient are presented in bold characters in the table.

None of the mathematics items, four of the reading items and two of the science items fell in this category.

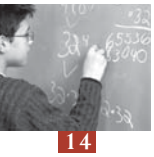


Table 14.10 ■ **Summary of item characteristics for the 2003 inter-country coder reliability (ICR) study**

Item	N	% agreement	% inconsistent	% too harsh	% too lenient
M150Q02	2457	95.7	2.0	0.7	1.6
M150Q03	2458	92.5	3.7	1.6	2.3
M155Q01	2456	98.7	0.4	0.8	0.1
M155Q02	2456	94.3	2.7	0.9	2.0
M155Q03	2458	92.9	3.0	1.6	2.5
M406Q01	2457	97.9	0.8	0.4	0.9
M406Q02	2457	96.6	1.7	0.2	1.6
M406Q03	2458	93.7	3.1	1.3	1.9
M413Q03	2456	89.7	3.7	1.6	4.9
M442Q02	2458	99.1	0.5	0.0	0.3
R055Q02	2460	89.8	3.5	1.8	4.8
R055Q03	2460	97.0	1.4	0.9	0.7
R055Q05	2460	97.0	1.2	1.1	0.7
R067Q04	2460	81.9	7.8	5.2	5.1
R067Q05	2460	84.5	5.9	4.8	4.8
R102Q04A	2460	98.7	0.6	0.3	0.4
R111Q02B	2460	77.1	7.8	3.0	12.1
R111Q06B	2460	85.0	5.4	1.9	7.7
R227Q03	2460	95.0	2.1	1.2	1.7
R227Q06	2460	99.3	0.3	0.2	0.3
S114Q03	2458	93.5	3.2	1.5	1.8
S114Q04	2457	80.1	7.7	3.6	8.6
S114Q05	2458	88.4	3.7	3.0	4.9
S129Q02	2457	94.0	2.2	1.4	2.4
S131Q02	2458	94.3	1.5	1.7	2.4
S131Q04	2458	89.9	3.2	4.0	3.0
S268Q02	2455	89.9	3.4	2.5	4.2
S326Q01	2456	92.0	3.2	3.4	1.4
S326Q02	2455	94.1	2.7	2.1	1.1
S327Q02	2458	84.7	5.1	1.6	8.6

Results by country

The materials explored in the ICR study included some of the most complex open-ended items used in PISA 2003. For many of the student answers that were flagged, perfect agreement was often difficult to reach, even between the consortium's adjudicators. In this context, as shown in Table 14.11, the proportion of cases when all 4 national coders and the international verifier gave exactly the same codes can be considered as reasonably high (on average, 79.7 per cent, with a standard deviation of 4.6). The lowest percentages of perfect agreement across all five coders were observed in Tunisia (72 per cent) and in New Zealand, Spain (Catalonian region) and Mexico (about 73 per cent).

The within-country rate of agreement was obviously somewhat higher (on average; 85.5 per cent of the cases received identical codes from the four national coders, with a standard deviation of 6.1). However, internal agreement was implausibly high in a few countries. Portugal, Indonesia, Turkey and Spain (Basque region) had 97 to 99 per cent agreement between the four national coders, that is, two standard deviations higher than the average rate across countries, suggesting that those countries may have implemented some undocumented procedure to increase the consistency of their multiple coding.

Table 14.11 ■ **Per cent of cases of perfect agreement between coders**

	N	Agreement between all national coders	Agreement between national coders and verifier
Australia	1800	84.7	82.3
Austria	1800	84.4	81.1
Belgium (French)	720	84.2	80.8
Belgium (Flemish)	1080	81.7	74.4
Canada (English)	900	82.4	77.4
Canada (French)	900	79.9	75.1
Czech Republic	1800	86.3	79.1
Denmark	1800	79.3	75.7
Finland	1800	91.5	86.8
France	1777	79.5	76.3
Germany	1800	82.4	80.1
Greece	1800	83.3	79.7
Hong Kong-China	1800	83.1	77.0
Hungary	1800	82.2	77.3
Indonesia	1790	98.2	82.5
Ireland	1800	79.0	75.9
Iceland	1800	87.1	80.1
Italy	1800	86.3	81.1
Japan	1800	95.7	88. 0
Korea	1800	87.1	81.7
Latvia (Latvian)	1200	83.6	77.2
Latvia (Russian)	600	84.7	78.5
Luxembourg (German)	1185	89.6	84.1
Luxembourg (French)	590	93.6	88.0
Mexico	1800	78.1	73.6
Netherlands	1800	81.2	76.1
Norway	1800	82.5	78.8
New Zealand	1800	74.8	72.7
Poland	1800	91.2	85.2
Portugal	1800	98.9	88.7
Russian Federation	1800	92.8	81.8
Scotland	1800	91.1	85.2
Serbia	1800	83.7	80.1
Slovak Republic	1800	81.3	75.8
Spain (Basque)	1800	97.2	86.7
Spain (Catalan)	1800	79.6	73.4
Spain (Castilian)	1799	80.2	74.0
Sweden	1800	82.8	78.4
Switzerland (German)	1260	79.9	76.8
Switzerland (French)	540	80.0	75.2
Thailand	1800	89.1	79.9
Tunisia	1800	76.8	71.9
Turkey	1800	97.1	88.8
United Kingdom	1800	90.7	85.4
United States	1800	87.3	82.0
Uruguay	1800	79.4	76.8
Mean		**85.3**	**79.7**
Std		**6.1**	**4.6**

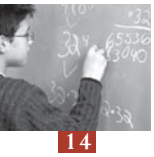


Table 14.12 ■ **ICR summary by countries and by domains**

		Mathematics			Science			Reading		
	N per domain	% agree-ment	% too harsh	% too lenient	% agree-ment	% too harsh	% too lenient	% agree-ment	% too harsh	% too lenient
Australia	600	96.7	0.5	1.0	90.0	1.3	4.7	93.3	0.8	2.0
Austria	600	95.0	1.0	1.5	90.8	3.5	1.5	90.2	1.7	2.7
Belgium	600	93.5	0.7	1.8	85.0	2.8	6.2	90.0	2.2	4.3
French	240	95.8	0.0	0.0	87.9	4.6	2.5	92.1	0.8	3.8
Flemish	360	**91.9**	1.1	3.1	**83.1**	1.7	**8.6**	88.6	3.1	4.7
Canada	600	95.3	1.0	0.9	89.1	1.8	4.8	91.1	0.6	4.6
English	300	96.0	1.0	1.0	89.0	2.0	4.7	91.0	0.3	4.7
French	300	94.0	1.0	0.7	89.3	1.3	5.0	91.3	1.0	4.3
Czech Republic	600	94.7	0.3	3.0	88.8	0.2	6.3	91.0	1.3	4.7
Denmark	600	95.7	0.7	1.3	94.5	0.5	1.7	88.5	2.5	3.0
Finland	600	96.3	1.3	1.2	93.2	1.7	3.5	93.3	2.2	1.7
France	587	93.7	0.7	1.9	90.5	1.9	3.1	88.5	2.3	3.3
Germany	600	94.8	0.5	1.3	91.3	1.3	1.7	92.5	0.8	1.7
Greece	600	95.3	0.0	1.5	96.8	0.8	2.2	89.7	0.8	4.0
Hong Kong-China	600	94.8	1.5	1.3	87.3	4.2	3.0	90.8	1.0	3.8
Hungary	600	94.7	1.0	1.7	91.5	4.7	1.8	87.3	**5.5**	2.0
Indonesia	600	98.3	0.8	0.8	91.7	2.7	4.9	**84.7**	**4.8**	**10.0**
Iceland	600	95.0	1.0	1.0	95.0	1.0	1.5	92.5	1.8	1.5
Ireland	600	93.8	0.8	2.5	85.7	1.7	5.5	91.0	1.8	1.8
Italy	600	95.5	0.8	1.5	93.3	1.3	1.8	92.3	2.0	1.5
Japan	600	97.7	1.7	0.5	90.7	2.8	6.2	91.8	0.8	5.5
Korea	600	95.8	1.0	1.8	90.8	3.8	3.2	91.3	0.8	4.5
Latvia	600	95.2	1.7	2.0	86.0	5.7	3.7	88.3	2.3	6.0
Latvian	400	95.3	1.5	2.0	85.3	**6.5**	3.3	87.0	3.0	6.0
Russian	200	95.0	2.0	2.0	87.5	4.0	4.5	91.0	1.0	6.0
Luxembourg	585	95.6	1.4	1.9	94.7	1.0	2.4	94.2	2.2	1.2
German	395	94.4	1.8	2.5	94.6	1.0	2.6	94.0	2.5	1.3
French	190	97.9	0.5	0.5	95.0	1.0	2.0	94.5	1.5	1.0
Mexico	600	97.3	0.2	0.8	87.0	1.7	4.8	88.0	1.7	4.2
Netherlands	600	93.5	**2.3**	1.3	87.2	4.3	1.5	89.2	2.2	3.8
New Zealand	600	95.0	0.7	1.7	85.3	2.0	2.7	91.3	2.2	1.7
Norway	600	95.2	0.7	1.3	90.5	1.2	3.8	91.5	1.3	3.3
Poland	600	95.7	0.0	2.7	93.0	2.5	3.0	95.0	0.8	3.2
Portugal	600	97.3	1.2	1.0	90.3	**7.0**	2.7	96.7	1.7	1.7
Russian Federation	600	93.8	0.8	1.5	89.7	**6.2**	2.8	87.2	**5.2**	**7.2**
Scotland	600	94.7	1.8	3.2	90.3	3.0	6.3	94.0	3.2	2.0
Serbia	600	96.5	0.3	1.2	88.7	3.8	3.0	91.3	3.3	3.8
Slovak Republic	600	95.0	0.5	2.0	86.8	1.3	**8.2**	**84.8**	2.5	5.2
Spain	1800	94.5	1.3	1.9	88.7	1.4	5.5	91.1	1.4	0.4
Basque	600	95.7	0.5	3.0	91.3	1.7	6.8	93.2	1.3	5.5
Catalan	600	**88.3**	0.3	**7.2**	85.3	1.8	6.3	**84.7**	2.7	6.8
Castilian	599	95.3	1.7	0.5	88.7	1.2	5.0	91.8	1.2	3.7
Sweden	600	94.2	0.3	1.8	88.5	0.8	**7.5**	90.8	0.3	4.3
Switzerland	600	93.0	1.3	3.3	90.2	1.7	3.5	88.5	1.5	4.3
German	420	**91.7**	1.7	4.0	90.2	1.9	3.1	89.5	1.0	4.5
French	180	96.1	0.6	1.7	90.0	1.1	4.4	86.1	2.8	3.9
Thailand	600	95.0	**2.2**	2.8	92.5	3.8	2.5	**83.2**	**6.0**	6.3
Tunisia	600	94.7	2.2	1.3	**84.7**	2.7	4.3	87.0	2.2	**7.2**
Turkey	600	96.3	1.2	2.5	93.0	3.3	2.3	92.5	1.0	6.0
United Kingdom	600	95.3	0.0	3.0	90.7	3.8	4.0	93.7	1.8	1.5
United States	600	95.0	0.7	1.0	91.2	2.3	4.2	92.8	1.7	3.7
Uruguay	600	95.8	0.3	1.0	91.8	0.8	2.5	91.8	2.0	2.0
All	**Mean**	**95.1**	**0.9**	**1.8**	**90.1**	**2.5**	**3.8**	**90.5**	**2.0**	**3.8**
	SD	**1.7**	**0.6**	**1.2**	**3.0**	**1.6**	**1.9**	**3.0**	**1.3**	**1.9**

© OECD 2005

Table 14.12 presents a summary of the ICR results by country and by domain. Multilingual countries are presented in the table with separate results by language, since independent teams coded the booklets for each language, and the quality of coding may have varied across the various groups of national coders. However, for reporting purposes, overall results were also computed by aggregating the data at the national level and weighting them according to the proportion of students assessed in each language.

Figures that differ from the overall percentages by more than one standard deviation appear in bold characters.

Only the Catalonian region of Spain seemed to show some slight systematic trend towards leniency in all 3 domains. In other countries, very few cases of biased or inconsistent codes occurred in mathematics, while science and reading appeared to be somewhat more problematic.

In science, some harshness was observed in Latvia (for the Latvian booklets), the Russian Federation and Portugal, while a relatively high percentages of lenient codes were observed in Belgium (for the Flemish booklets), the Slovak Republic and Sweden.

In reading, three countries tended to have both too harsh and too lenient codes for a number of items (Indonesia, the Russian Federation and Thailand). Lenient codes were observed in Tunisia and harsh codes in Hungary.

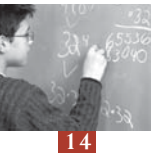
Poorly coded items by country

In fact, most of the coding problems observed in the various countries tended to affect particular items, rather than having a more general effect. This suggests that either the scoring instructions or the training of national coders could be improved for specific items. Table 14.13 lists items that appeared to have relatively serious problems in specific countries, that is, one or more of the following:

- Items with too low within-country agreement (less than 60 per cent identical marks across the four national coders);
- Items with too low international agreement (where the agreement between the codes given by the majority of national coders and the verifier or the adjudicators was less than 70 per cent);
- Items with 20 per cent or more cases where the codes given by the national coders were found to be too harsh;
- Items with 20 per cent or more cases where the codes given by the national coders were found to be too lenient;
- Items with more than 20 per cent of cases when the codes given by the national coders were too inconsistent to be compared with the verifier's code.

An analysis was conducted to investigate whether the items with coding problems in a specific country showed particularly poor statistics for that country in the item analysis (in terms of item fit and differential item functioning).

Many items for which the inter-country reliability study evidenced problems in the national coding proved to have large fit indices. The correlation between fit index and per cent of agreement between national codes and codes given by the verifier or adjudicator was -0.32



(N = 30 items × 46 national samples = 1 380 cases). Moderately high correlations were also observed between percentages of too harsh or too lenient codes and item fit (respectively, 0.22 and 0.18).

One might also expect a significant correlation between the percentage of too harsh or too lenient codes awarded by national coders and possible differential item functioning. Actually the correlations, though significant, were quite low ($r = 0.12$ between per cent too harsh and high delta; and $r = 0.14$ between per cent too lenient and low delta). This may be due to the fact that in many countries, when an item showed a high percentage of too lenient codes, it often also showed non-negligible occurrences of too harsh codes.

Table 14.13 ■ **Poorly coded items by country**

	Item	N	Inter-national agreement	Too inconsistent	Too harsh	Too lenient	All national coders agree
AUS	S114Q04	60	65.0	15.0	3.3	16.7	71.7
BEF	S129Q02	24	75.0	4.2	20.8	0.0	91.7
BEN	R111Q02B	36	58.3	8.3	2.8	30.6	52.8
BEN	S327Q02	36	36.1	13.9	0.0	50.0	58.3
CAF	R111Q02B	30	66.7	6.7	3.3	23.3	53.3
CHD	R111Q06B	42	66.7	4.8	2.4	26.2	61.9
CHF	R067Q04	18	72.2	22.2	5.6	0.0	66.7
CHF	R111Q02B	18	55.6	22.2	0.0	22.2	50.0
CHF	S327Q02	18	61.1	11.1	0.0	27.8	61.1
CZE	S327Q02	60	56.7	28.3	0.0	15.0	58.3
ESB	S327Q02	60	75.0	0.0	0.0	25.0	100.0
ESC	R067Q05	60	66.7	16.7	11.7	5.0	60.0
ESC	R111Q02B	60	56.7	15.0	5.0	23.3	40.0
ESC	R111Q06B	60	68.3	6.7	5.0	20.0	68.3
ESC	S114Q04	60	58.3	8.3	0.0	33.3	63.3
FRA	S114Q04	59	67.8	13.6	8.5	10.2	35.6
HKG	R111Q06B	60	73.3	5.0	0.0	21.7	78.3
IRL	S327Q02	60	56.7	25.0	0.0	18.3	36.7
JPN	R111Q02B	60	61.7	6.7	1.7	30.0	68.3
JPN	S327Q02	60	58.3	0.0	0.0	41.7	91.7
KOR	S327Q02	60	73.3	1.7	0.0	25.0	81.7
NZL	S114Q04	60	60.0	33.3	5.0	1.7	31.7
PRT	S326Q01	60	60.0	0.0	38.3	1.7	95.0
RUS	R111Q02B	60	58.3	0.0	3.3	38.3	98.3
RUS	S131Q02	60	70.0	0.0	28.3	1.7	100.0
SVK	R067Q04	60	65.0	15.0	8.3	11.7	36.7
SVK	R111Q02B	60	66.7	20.0	1.7	11.7	46.7
SVK	S114Q04	60	61.7	10.0	0.0	28.3	46.7
SWE	R111Q02B	60	65.0	13.3	0.0	21.7	36.7
SWE	S327Q02	60	65.0	5.0	0.0	30.0	70.0
THA	R067Q04	60	56.7	11.7	23.3	8.3	68.3
THA	R067Q05	60	60.0	8.3	16.7	15.0	70.0
THA	R111Q02B	60	68.3	8.3	5.0	18.3	68.3
TUN	R111Q02B	60	68.3	6.7	0.0	25.0	60.0

Note

1 These figures represent the materials received from all PISA 2003 participating countries, except Brazil. The Brazilian ICR booklets were received too late to be included in the study.

15

Data Adjudication

INTRODUCTION

This chapter describes the process used to adjudicate the implementation of PISA 2003 in each of the participating countries, and gives the outcomes of the data adjudication. In particular, this chapter reviews the:

- Extent to which each country met PISA sampling standards;
- Outcomes of the national centre and PISA quality monitoring visits;
- Quality and completeness of the submitted data;
- Outcomes of the inter-country reliability study; and
- Outcomes of the translation verification process.

The standards for PISA 2003, which were formally presented to the National Project Managers (NPMs) at the Brussels NPM meeting in February 2001, were used as the basis for the adjudication. The latest version of the standards is available on the PISA Web site (*www.pisa.oecd.org*). The issues covered in those standards are:

- Sampling
- Translation and verification
    - Selection of translators
    - Submission of questionnaire adaptations and modifications for approval
    - Submission of material for translation and verification
- Test administration
    - Selection of test administrators
    - Training of test administrators
    - Security of material
    - Testing session
- Quality monitoring
    - Site visits and training of PISA quality monitors (PQM)
    - Visit by PISA quality monitors
- Coding
    - Single coding
    - Multiple coding
    - PISA international standard indicators [Inter-country-rater-reliability study]
- Data entry and submission
    - Materials submitted
    - Data cleaning

Implementing the standards – quality management

NPMs of countries and adjudicated regions were responsible for implementing the standards based on consortium advice as contained in the study's various operational manuals. During the implementation phase the consortium conducted two quality management activities. The first was quality control performed by consortium staff as they worked with NPMs to implement key parts of the project. As part of the quality control activities, consortium staff checked the work of NPMs and provided advice on rectifying action when required and before critical errors occurred. The second was quality monitoring, which involved the systematic collection of data that monitored the implementation of the standards. For data adjudication it was the information collected during both the quality control and quality monitoring activities that was used to determine the level of compliance to the standards.

Information available for adjudication

The information collected by consortium staff during their quality control activities included communications and documentation exchanged with NPMs. The information available from quality monitoring instruments included:

- PISA quality monitor reports (data collection sheets and general observations);
- Test administrator session reports;
- Main study reviews;
- Sampling forms;
- National centre quality monitor interviews; and
- Data cleaning questionnaire.

Each of the quality monitoring instruments addressed different aspects of the standards and were collected at different times during the data collection phase. There are two types of PISA Quality Monitor (PQM) reports, one containing data for each observed session and another detailing the general observations of each quality monitor. The PQM reports contain data related to test administration as well as a record of interview with school co-ordinators. The test administrator session report is completed by each test administrator after each test session and also contains data related to test administration. The data from this report were data-entered by the national centre and submitted as part of the dataset to the consortium. The national centre quality interview schedule contains information on all the standards, as does the main study review. The data submission questionnaire contains information specific to the data and is mainly used for data cleaning purposes.

The national centre quality monitor interview schedule, main study review, and data submission questionnaire are self-declared by the NPM. The PQM data is collected independently of the NPM and can be viewed as being collected by a peer of the test administrator who is nominated by the NPM.

Data adjudication process

The main aim of the adjudication process is to make a single determination on adjudicated data in a manner that is transparent, based on evidence and which is defensible. The data adjudication process achieved this through the following steps:

- *Step 1*: Quality control and quality monitoring data were collected during the data collection phase.

- *Step 2:* Data from quality monitoring instruments were entered into a single quality management database.
- *Step 3:* Experts compiled country-by-country reports that contained quality monitoring data for expert areas.
- *Step 4:* Experts considered the quality monitoring data, along with their quality control information, in order to make a judgement. In this phase the experts collaborated with the project director and data manager to address any identified areas of concern. Where necessary, the relevant NPM was contacted through the project director. At the end of this phase each expert constructed, for each adjudicated-dataset, a summary detailing how the standards had been implemented.
- *Step 5:* The consortium reviewed the reports and made a determination with regard the quality of the data.

It was expected that the data adjudication would result in a range of possible recommendations. Some possible, foreseen recommendations included:

- That some data be removed for a particular country, for example the removal of data for some items, such as open-ended items, or the removal of data for some schools.
- That rectifying action be performed by the NPM, for example providing additional evidence to demonstrate that there is no non-response bias or rescoring open-ended items.
- That the data not be endorsed for use in certain types of analyses.
- That the data not be endorsed for inclusion in the PISA 2003 database.

Throughout the data collection phase, the consortium concentrated its quality control activities to ensure that the highest scientific standards were implemented. However during data adjudication a wider definition of quality was used especially when considering data that was at risk. In particular the underlying criteria used in adjudication was "fitness for use". That is, data was endorsed for use if it was deemed to be fit for meeting the intended purposes of PISA 2003.

GENERAL OUTCOMES

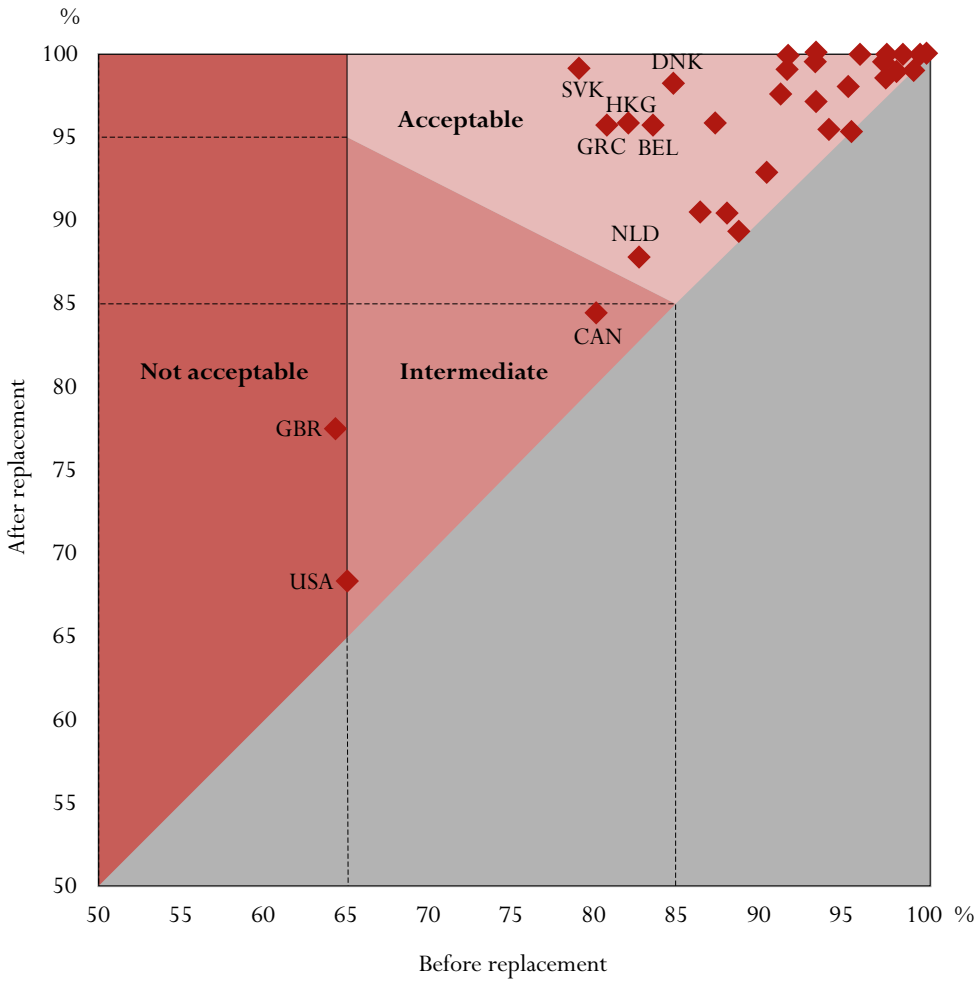
Overview of response rate issues

The PISA school response rate requirements are discussed in Chapter 4. Figure 15.1 is a scatter plot of the attained PISA school response rates before and after replacements. Those countries that are plotted in the lighter shaded region were regarded as fully satisfying the PISA school response rate criterion.

Canada, United Kingdom, and the United States failed to meet the school response rate requirements. In addition to failing the school response rate requirement, the United Kingdom was the only participant to fail the student response rate requirement (see Table 12.4).

After reviewing the sampling outcomes, the consortium asked Canada, United Kingdom, and The United States, to provide additional data that would assist the consortium in making a balanced judgement about the threat of the non-response to the accuracy of inferences which could be made from the PISA data.

Figure 15.1 ■ **Attained school response rates**



DETAILED COUNTRY COMMENTS

It is important to recognise that the PISA data adjudication is a late but not necessarily final step in a quality assurance process. By the time each country was adjudicated, quality assurance mechanisms (such as the sampling procedures documentation, translation verification, data cleaning and site visits) had identified a range of issues and ensured that they had been rectified at least in the majority of cases. Details on the various quality assurance procedures and their outcomes are documented elsewhere (see Chapter 7 and Appendix 9). Data adjudication focused on residual issues that remained after these quality assurance processes. There were not many such issues and their projected impact on the validity of the PISA results was deemed to be negligible. Unlike sampling issues, which under most circumstances could directly affect all of a country's data, the residual issues identified in other areas have an impact on only a small proportion of the data. For example, coding leniency or severity for a single item in reading has an effect on between just one-third and one half of 1 per cent of the reading data and even for that small fraction, the effect would be minor. Other breaches of standards identified in a small number of countries include

a failure to follow the specified multiple marker design and a failure to involve national committees in instrument development. Where the specified multiple coding design was not implemented, a sufficient level of quality assurance data was usually available to determine the quality of the manual coding.

Australia

Australia fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Austria

Austria fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Belgium

Belgium fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Brazil

Brazil had a very low coverage of the 15-year-old population (54 per cent), due to low rates of enrolment, and that this should be taken into account when interpreting Brazilian data. Further, Brazil did not submit data for the scheduled inter-country coder reliability study and therefore it was not possible to implement the necessary quality assurance procedures for the manually coded items.

The Brazilian data was available for inclusion in the full range of PISA reports.

Canada

There were sampling-related concerns with the Canadian data. The overall exclusion rate of 6.83 per cent exceeded the PISA standard of 5 per cent. The majority of the exclusions (5.26 per cent) were within-school exclusions with large contributions from language-based exclusions and special needs students. The high overall exclusion rate was also contributed to by the exclusion of very small schools, that is, schools having only one or two eligible students. In addition there was also a high ineligible rate of 5.29 per cent, where the ineligibles were about evenly split between drop-outs and transferred students.

The Canadian school response rate, of 79.95 per cent before replacement and 84.38 per cent after all replacements, did not meet PISA standards. Much of Canada's non-responses came from the relatively large province of Ontario. Canada presented evidence to show that the characteristics of non-responding schools in Ontario were not markedly different from those of respondent schools.

It was concluded that the problems observed in the Canadian data had a minimal impact on the data, and inclusion in the full range of PISA 2003 reports was recommended.

Czech Republic

The Czech Republic fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Denmark

Denmark had an overall exclusion rate of 5.33 per cent, the majority of which were within school exclusions due to language issues. This exceeds the PISA standard of 5 per cent.

Inclusion of Danish data in the full range of PISA 2003 reports was recommended.

Finland

Finland fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

France

The implementation of PISA in France deviated from the internationally recommended procedures in a number of ways. First, France did not implement the school questionnaire. It follows that France cannot be included in those reports and analyses that utilise school questionnaire data. Second, France did not implement the recommended multiple coding design. The alternative design implemented in France, however, was carefully reviewed and it was deemed that the design implemented provided a sufficient level of quality assurance for the coding activities. Third, it was noted that the test administrators were not trained in person as required by the standards. As an alternative, the test administrators were trained through phone calls. Finally, due to local requirements, the PQMs were school inspectors and were not formally independent of the French national centre as was required by the standards.

Given that the PISA quality monitors did not identify problems with the test administration and that the lack of independence of the quality monitors was unlikely to cause problems it was concluded that the identified issues would have no marked effect on the data and it was therefore recommended that all the available French data be included in PISA reports.

Germany

Germany fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Greece

Problems were identified with the printing and pagination of the instruments administered in Greece. Additional analysis undertaken to examine this issue suggested that at the national level the impact of printing problems on the data were likely to be minimal. It was recommended that the Greek data be included in the full range of PISA 2003 reports.

Hong Kong-China

Hong Kong-China fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Hungary

Hungary fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Iceland

Iceland fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Indonesia

Indonesia did not submit the educational career questionnaire for international verification before the questionnaire was administered in the field. Also, there was evidence of poor translation in some of the administered instruments. The consortium therefore deleted, during the analysis phase, items it identified as poorly translated (see Chapter 5). The quality of the printed instruments was also significantly below that of other PISA countries. While coverage of the PISA population met PISA standards, Indonesia had a low level of 15-year-old enrolment, so coverage of 15-year-olds was just 46 per cent.

It was recommended that all the available Indonesian data be included in PISA 2003 reports.

Ireland

Ireland fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Italy

Italy fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Italy, Veneto - Nord Est

The Italian region of Veneto - Nord Est fully met the PISA 2003 standards.

Italy, Trento - Nord Est

The Italian region of Trento - Nord Est fully met the PISA 2003 standards.

Italy, Toscana – Centro

The Italian region of Toscana – Centro fully met the PISA 2003 standards.

Italy, Piemonte - Nord Ovest

The Italian region of Piemonte - Nord Ovest fully met the PISA 2003 standards

Italy, Lombardia - Nord Ovest

The Italian region of Lombardia - Nord Ovest fully met the PISA 2003 standards.

Italy, Bolzano - Nord Est

The Italian region of Bolzano - Nord Est fully met the PISA 2003 standards.

Japan

Japan fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Latvia

Latvia did not submit the Russian language test instruments to the international verification team for a final optical check. In addition the Russian coding guides were not submitted for verification. In Latvia, 35.4 per cent of the population is assessed in Russian. Analysis of the submitted data suggested that these breaches of PISA 2003 standards had no marked affect on the Latvian data and inclusion in the full range of PISA 2003 reports was recommended.

Liechtenstein

Liechtenstein fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Luxembourg

In Luxembourg, students were tested in either German or French, depending upon the combination of languages in which they have experienced instruction. The procedures of allocating languages to students were different in PISA 2003 to those applied in PISA 2000. This change in procedures was deemed to prevent the interpretation of trends in Luxembourg between PISA 2000 and PISA 2003.

Luxembourg fully met the PISA 2003 standards, and inclusion in PISA 2003 reports that were not concerned with trends was recommended.

Mexico

For Mexico, it was noted that the quality of the printing and layout of instruments varied in the administered booklets. The originally submitted database included unusually high numbers of inconsistencies between student questionnaire data and tracking forms, which could only be corrected by taking the information provided in the tracking forms as accurate. Some school questionnaire indicators were found to have percentages of missing values around 50 per cent after data cleaning. Consequently, some of these indicators were not included in the final database.

Furthermore, the percentage of ineligible students was very high (8.10 per cent), and this was due mainly to a substantial number of students with invalid or out-of-range incorrect birth dates, and transferred students. The coverage of the national 15-year-old population was low (49 per cent), primarily because of low (58 per cent) enrolment rates of the target population. As the problems encountered with sampling and data collection were not deemed to have marked effects on the results, inclusion in the full range of PISA 2003 reports was recommended.

The Netherlands

The Netherlands fully met the PISA standards, and inclusion in the full range of PISA 2003 reports was recommended.

New Zealand

The within-school samples included a high percentage of ineligible students (5.99 per cent), with these approximately evenly split between drop-outs and transferred students. Additionally, New Zealand had an overall exclusion rate of 5.07 per cent, the majority of which were within-school exclusions due to language issues.

It was recommended that the data for New Zealand be included in the full range of PISA 2003 reports.

Norway

Norway fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Poland

Norway fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Portugal

In Portugal, the within-school samples included a high percentage of ineligible students (5.68 per cent), mostly being due to dropouts. It was recommended that the data for Portugal be included in the full range of PISA 2003 reports.

Korea

Korea fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Russian Federation

The Russian Federation fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Serbia

In Serbia, a number of sampling-related concerns were noted. First, the overall exclusion rate of 5.66 per cent, does not meet the PISA standard of 5 per cent. Second, the within-school samples included a high percentage of ineligible students (5.74 per cent), with those being mostly drop-outs. Third, while coverage of the PISA population met PISA standards, Serbia had some frame data issues so coverage of 15-year-olds appeared to be just 69 per cent.

In addition, Serbia implemented an unapproved marker design. Follow-up analysis suggested that this had no marked effect on the data.

It was recommended that the Serbian data be included in the full range of PISA 2003 reports.

Spain

It was noted that Spain had high overall exclusion rates (7.29 per cent) that did not meet PISA 2003 standards. This high level of exclusions was largely due to within-school exclusions. Additionally, the third

coverage index was low but has been explained by sources of error in the statistics gathered to obtain the SF2[a] value.

In the Basque country, as noted below, some students were tested in a language spoken at home rather than the official language of instruction. However, the percentage affected in Spain as a whole was very small.

It was recommended that the Spanish data be included in the full range of PISA reports.

Castilla-Leon

The Spanish region of Castilla-Leon had an overall exclusion rate (5.96 per cent, but 4.89 per cent when languge exclusions were removed), and inclusion in the full range of PISA reports was recommended.

Catalonia

The Catalonian multiple-marker data showed a consistent leniency bias across all three domains included in the study. The impact on the overall results, however, was deemed to be small. It was concluded that the Spanish region of Catalonia fully met the PISA standards.

Basque Country

For the Spanish region of the Basque Country, the standard procedure relating to the language of assessment was not followed. All students receiving instruction in bilingual Spanish/Basque settings were tested in Castillian, instead of being given the choice of a Basque or Spanish booklet. Students receiving instruction in Basque immersion schools were only tested in Basque when they had a Basque-speaking mother, a Basque-speaking father and used themselves Basque in their communications at home. All other Basque immersion students were tested in their home language (Castillian) rather than in their language of instruction (Basque). Note that as the Basque Country contains only a small percentage of the Spanish population this deviation does not influence the results for Spain overall.

In all other respects the data for the Basque Country met the PISA standards. The consortium recommended that the Basque Country data be included in the full range of PISA reports and that the data be annotated where it is published to indicate that the PISA results in the Basque Country must be interpreted as the results obtained by the students enrolled in the Basque educational system, but not as the results obtained by the students attending instruction in Basque language.

Slovak Republic

The Slovak Republic fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Sweden

Sweden fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Switzerland

Switzerland fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Thailand

Thailand fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Tunisia

The within-school samples included a high percentage of ineligible students (6.36 per cent). Further, it was noted that the printing quality of the administered instruments varied, and that there were pagination and layout errors in some of the administered booklets. Follow-up analysis did not suggest that the low printing quality had had a material affect on the quality of the data.

Tunisia met the PISA standards, and inclusion in the full range of PISA reports was recommended.

Turkey

While coverage of the PISA population met PISA 2003 standards, Turkey had a low level of 15-year-old enrolment (54 per cent) so coverage of 15-year-olds was just 36 per cent. Turkey has several forms of informal education through which participants receive their training via mail, television, or hands on experience. There are no records of the 15-year-olds that might be in those programmes. This may be one factor explaining the low enrolment in formal education of 15-year-olds.

Turkey fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

United Kingdom

Problems relating to response rate and testing window were identified for the data from the United Kingdom. A poor school response rate resulted in an extension of the three-month testing window, which is required by the PISA technical standards. After the extension of the testing window, the school response rate (64.32 per cent prior to replacements and 77.37 per cent after replacements) and student response rates (77.92 per cent) were still below PISA standards.

The United Kingdom was especially well placed to provide accurate evidence one way or the other as to the existence of non-response bias in the PISA data, because results of national assessment data were available at the school level (for two assessments) and at the individual student level (for one of these assessments) for the entire PISA sample. The United Kingdom national centre prepared a report in February 2004, entitled *PISA 2003 England Sample: Report of an Investigation into Response Bias at the School and Student Level*. While England and Wales were part of the same data collection procedure, data from Scotland, which was adjudicated by the consortium as a separate unit, were fully comparable with results from other OECD countries and with results from PISA 2000.

The conclusion that the PISA sampling referee drew from this analyses was that there was good evidence that the school sample was not substantially biased upwards or downwards, in terms of mean student achievement, as a result of non-response. However, there was evidence that the responding schools were a more homogeneous group in terms of student achievement than the full sample.

For the student sample, the conclusion was that it appeared that student non-response was likely to have induced a bias in achievement. It was not possible to ascertain the exact magnitude of this. However, before

finalising this conclusion, an important additional check was needed. The initial analyses on response rates were carried out before student weights had been calculated by the consortium (see Chapter 8 for a full description of student weights). The PISA sampling referee, therefore, asked the United Kingdom national centre to carry out analyses using the student weights with adjustments for non-response, to see whether these adjustments might have been effective in reducing the non-response bias. These weighted analyses indicated that the weight adjustments did not have an appreciable effect on reducing the non-response bias.

The uncertainties surrounding the sample and its bias are such that PISA 2003 scores for the United Kingdom cannot reliably be compared with those of other countries. They can also not be compared with the performance scores for the United Kingdom from PISA 2000. The regional data from Wales are also not comparable with other countries.

The results are, however, accurate for many within-country comparisons between subgroups (*e.g.* males and females) and for relational analyses. The results for the United Kingdom were included in a separate category below the results for the other participating countries. Other data for the United Kingdom that were not reported in the initial report were made available on the PISA Web site (*www.pisa.oecd.org*).

All international averages and aggregate statistics include the data for the United Kingdom.

Scotland

Scotland fully met the PISA standards.

United States

Problems relating to response rate and testing window were identified for the data from the United States. As a result of a poor school response rate the consortium approved the use of a second testing window. Both the use of a second testing window and the timing of the window within the school year were breaches of the PISA technical standards. After taking into account the data from the second testing window, the United States data still did not meet the school response standards, the overall school response rate was 64.94 per cent before replacements and 68.12 per cent after replacements. Furthermore, the United States had high overall exclusion rates (7.28 per cent) mostly due to high within school exclusions. These did not seem to be concentrated in any particular category of student (*i.e.* gender, grade, etc.) but were spread over all student types.

Two separate investigations were conducted to validate the United States data. The first investigated the hypothesis that testing students early in the school year would lead to different achievement results than testing students of equivalent age later in the school year, as PISA requires. The hypothesis is that, because of loss of retention of knowledge and skills over the summer period, a student of a particular age (at the time of testing), tested at the beginning of the school year will tend to perform less well on PISA than a student of the same age (relative to the testing date) tested at a later point in the year. Specifically for the United States, this would mean that students born between July 1987 and June 1988, tested in September and October 2003, would not perform as well on average as students born in 1987 and tested in April and May 2003.

Schools were not randomised to testing periods, but rather the time of testing is confounded with the school's willingness to participate in the April-May period. This willingness to participate at this time of

the year might well be associated with student achievement. The United States national centre conducted an analysis that attempted to deal with this issue of confounding. Although no nonrandomised study can ever be entirely conclusive, the evidence was quite strong that use of a later testing time did not impact the average achievement results, either negatively (as hypothesised above) or positively. The multi-level models used showed that while public/private status, school size, percent of minority students, location, and region all had significant relationships with student achievement, time of testing did not. It was also the case that the mean scores of students tested in September and October were almost identical to those tested in April and May, suggesting that this finding is robust (in other words, it is not necessary to rely on model to explain away any differences between the two time periods).

The second study was a non-response bias analysis, conducted on the assumption that the September and October assessments would in fact be included in the data. These analyses were conducted by the United States national centre. The PISA sampling referee reviewed this report and concluded that, there is likely to be relatively little school non-response bias. Region did appear to be significantly related to school response, but it was not a very strong predictor of achievement. It also appeared that the respondent sample was somewhat relatively deficient in Asian and Pacific Islander students. However, the absolute difference in the percentages of these students between the responding sample and full sample is not great (4.4 per cent in the full sample; 3.8 per cent in the responding sample).

The United States data was included in the full range of PISA 2003 reports.

Uruguay

While coverage of the PISA population met PISA standards, Uruguay had a low level of 15-year-old enrolment, so coverage of 15-year-olds was just 63 per cent. It was also noted that the percentage of ineligible students was high (7.78 per cent).

Uruguay met the PISA standards, and inclusion in the full range of PISA reports was recommended.

16

Proficiency Scale Construction

INTRODUCTION

The PISA test design makes it possible to use techniques of modern item response modelling (sometimes referred to as item response theory, or IRT) to simultaneously estimate the ability of all students taking the PISA assessment, and the difficulty of all PISA items, locating these estimates of student ability and item difficulty on a single continuum.

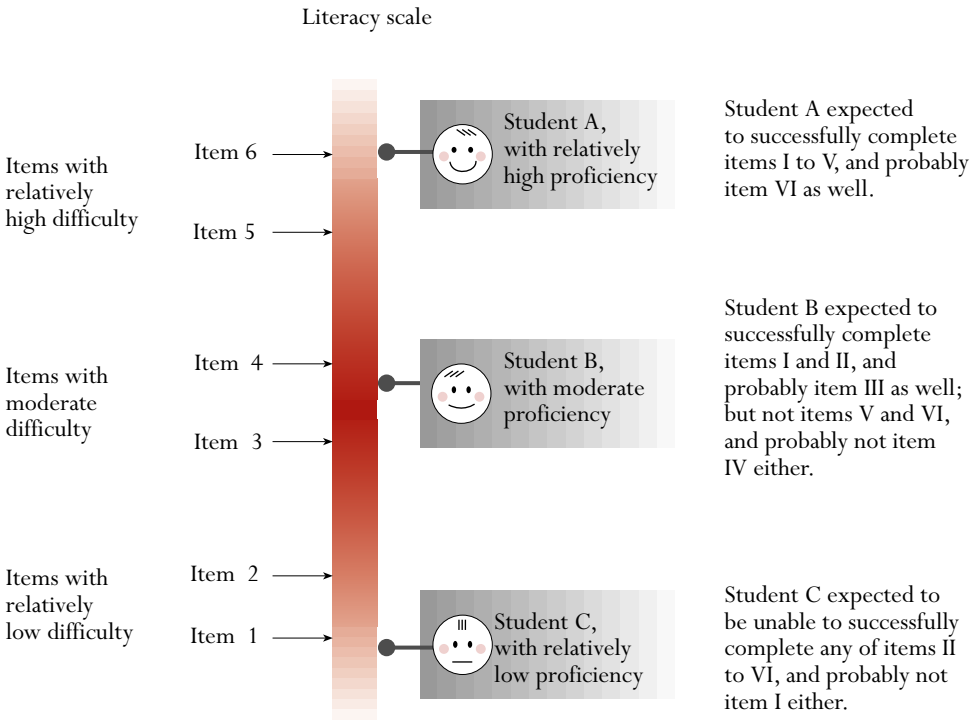
The relative ability of students taking a particular test can be estimated by considering the proportion of test items they get correct. The relative difficulty of items in a test can be estimated by considering the proportion of test takers getting each item correct. The mathematical model employed to analyse PISA data, generated from a rotated test design in which students take different but overlapping tasks, is implemented through test analysis software that uses iterative procedures to simultaneously estimate the likelihood that a particular person will respond correctly to a given test item, and the likelihood that a particular test item will be answered correctly by a given student. The result of these procedures is a set of estimates that enables a continuum to be defined, which is a realisation of the variable of interest. On that continuum it is possible to estimate the location of individual students, thereby seeing how much of the literacy variable they demonstrate, and it is possible to estimate the location of individual test items, thereby seeing how much of the literacy variable each item embodies. This continuum is referred to as the PISA literacy scale in the test domain of interest.

PISA assesses students and uses the outcomes of that assessment to produce estimates of students' proficiency in relation to a number of literacy variables. These variables are defined in the relevant PISA literacy framework. For each of these literacy variables, one or more scales are defined, which stretch from very low levels of literacy through to very high levels. When thinking about what such a scale means about student proficiency, it can be observed that a student whose ability estimate places them at a certain point on the PISA literacy scale would most likely be able to successfully complete tasks at or below that location, and increasingly more likely to complete tasks located at progressively lower points on the scale, but would be less likely to be able to complete tasks above that point, and increasingly less likely to complete tasks located at progressively higher points on the scale. Figure 16.1 depicts a literacy scale, stretching from relatively low levels of literacy at the bottom of the figure, to relatively high levels towards the top. Six items of varying difficulty are placed along the scale, as are three students of varying ability. The relationship between the students and items at various levels is described.

It is possible to describe the scales using words that encapsulate various demonstrated competencies typical of students possessing varying amounts of the underlying literacy constructs. Each student's location on those scales is estimated, and those location estimates are then aggregated in various ways to generate and report useful information about the literacy levels of 15-year-old students within and among participating countries.

Development of a method for describing proficiency in PISA reading, mathematical and scientific literacy occurred in the lead-up to the reporting of outcomes of the PISA 2000 survey. Essentially the same methodology was used again to develop proficiency descriptions for PISA 2003. Given the volume and breadth of data that were available from the PISA 2003 assessment, development of more detailed descriptions of mathematical literacy became possible. Proficiency descriptions were also newly developed for problem-solving skills. The detailed proficiency descriptions that had been developed for the reading domain in PISA 2000 were used again with the reduced data available from PISA 2003. The summary descriptions used for science in 2000 were used again in 2003.

Figure 16.1 ■ **The relationship between items and students on a proficiency scale**



The mathematics expert group and the problem-solving expert group worked with the consortium to develop sets of described proficiency scales for PISA mathematics and problem solving. Consultations regarding these described scales with the PISA Governing Board (PGB), the mathematics forum, National Project Managers (NPMs) and the PISA technical advisory group took place over several stages before their final adoption by the PGB.

This chapter discusses the methodology used to develop those scales and to describe a number of levels of proficiency in the different PISA literacy variables, and presents the outcomes of that development process.

DEVELOPMENT OF THE DESCRIBED SCALES

The development of described proficiency scales for PISA 2003 was carried out through a process involving a number of stages. The stages are described here in a linear fashion, but in reality the development process involved some backwards and forwards movement where stages were revisited and descriptions were progressively refined.

Stage 1: Identifying possible subscales

The first stage in the process involved the experts in each domain articulating possible reporting scales (dimensions) for the domain. For reading in the PISA 2000 survey cycle, two main options were actively considered – scales based on the type of reading task and scales based on the form of reading material. For the international report, the first of these was implemented, leading to the development of a scale for "retrieving information", a second scale for "interpreting texts" and a third for "reflection and evaluation".[1]

In the case of mathematics, a single proficiency scale was developed for PISA 2000, but with the additional data available in the 2003 survey cycle, when mathematics was the major test domain, the possibility of reporting according to the four overarching ideas or the three competency clusters described in the PISA mathematics framework were both considered. For science, a single overall proficiency scale was developed for the 2000 survey cycle, and this was again used to report results from PISA 2003. There had been interest in considering two subscales ("scientific knowledge" and "scientific processes"), but the small number of science items in PISA 2000 and PISA 2003, when science was a minor domain, meant that this was not possible. For PISA 2006, when science will be the major test domain, this matter will be revisited. For problem solving, similarly, a single scale was developed and described to the extent possible given the relatively small number of problem-solving items included in the survey.

Wherever multiple scales were under consideration, they arose clearly from the framework for the domain, they were seen to be meaningful and potentially useful for feedback and reporting purposes, and they needed to be defensible with respect to their measurement properties. Because of the longitudinal nature of the PISA project, the decision about the number and nature of reporting scales had to take into account the fact that in some test cycles a domain will be treated as minor and in other cycles as major. The amount of data available to support the development and application of described proficiency scales will vary from cycle to cycle for each domain, but the PGB expects that the consortium will develop proficiency scales that can be compared across survey cycles.

Stage 2: Assigning items to scales

The second stage in the process was to associate each test item used in the study with each of the scales under consideration. Mathematics experts (including members of the expert group, the test developers and consortium staff) judged the characteristics of each test item against the relevant framework categories. Later, statistical analysis of item scores from the field trial was used to obtain a more objective measure of fit of each item to its assigned scale.

Stage 3: Skills audit

The next stage involved a detailed expert analysis of each item, and in the case of items with partial credit, for each score step within the item, in relation to the definition of the relevant subscale from the domain framework. The skills and knowledge required to achieve each score step were identified and described.

This stage involved negotiation and discussion among the experts involved, circulation of draft material, and progressive refinement of drafts on the basis of expert input and feedback.

Stage 4: Analysing field trial data

For each set of scales being considered, the field trial item data were analysed using item response techniques to derive difficulty estimates for each achievement threshold for each item in each subscale.

Many items had a single achievement threshold (associated with getting the item right rather than wrong). Where partial credit was available, more than one achievement threshold could be calculated (achieving a score of one or more rather than zero, two or more rather than one, etc.).

Within each subscale, achievement thresholds were placed along a difficulty continuum linked directly to student abilities. This analysis gives an indication of the utility of each scale from a measurement perspective.

Stage 5: Defining the dimensions

The information from the domain-specific expert analysis (Stage 3) and the statistical analysis (Stage 4) was combined. For each set of scales being considered, the item score steps were ordered according to the size of their associated thresholds and then linked with the descriptions of associated knowledge and skills, giving a hierarchy of knowledge and skills that defined the dimension. Natural clusters of skills were found using this approach, which provided a basis for understanding each dimension and describing proficiency in different regions of the scale.

Stage 6: Revising and refining with main study data

When the main study data became available, the information arising from the statistical analysis about the relative difficulty of item thresholds was updated. This enabled a review and revision of Stage 5 by the working groups, and other interested parties. The preliminary descriptions and levels were then reviewed and revised in the light of further technical information that was provided by the technical advisory group, and the approach to defining levels and associating students with those levels that had been used in the reporting of PISA 2000 results was applied.

Stage 7: Validating

Two major approaches to validation were then considered, and used to varying degrees, by the mathematics and problem-solving working groups. One method was to provide knowledgeable experts (*e.g.* teachers, or members of the subject matter expert groups) with material that enabled them to judge PISA items against the described levels, or against a set of indicators that underpinned the described levels. Some use of such a process was made, and further validation exercises of this kind may be used in the future. Second, the described scales were subjected to an extensive consultation process involving all PISA countries through their NPMs. This approach to validation rests on the extent to which users of the described scales find them informative.

DEFINING PROFICIENCY LEVELS

How should the proficiency continuum be divided into levels that might have some utility? And having defined levels, how should the level to which a particular student should be assigned be decided? What does it mean to "be at a level"? The relationship between the student and the items is probabilistic – there is some probability that a particular student can correctly do any particular item. If a student is located at a point above an item, the probability that the student can successfully complete that item is relatively high, and if the student is located below the item, the probability of success for that student on that item is relatively low.

This leads to the question as to the precise criterion that should be used in order to locate a student on the same scale on which the items are laid out. When placing a student at a particular point on the scale, what probability of success should be insisted on in relation to items located at the same point on the scale? If a student were given a test comprising a large number of items each with the same specified difficulty, what proportion of those items could the student be expected to successfully complete? Or, thinking of it in another way, if a large number of students of equal ability were given a single test item with a specified item difficulty, about how many of those students would be expected to successfully complete the item?

The answer to these questions is essentially arbitrary, but in order to define and report PISA outcomes in a consistent manner, an approach to defining performance levels, and to associating students with those levels, is needed. The methodology that was developed and used for PISA 2000 will be essentially retained for PISA 2003.

Defining proficiency levels for PISA 2000 progressed in two broad phases. The first, which came after the development of the described scales, was based on a substantive analysis of PISA items in relation to the aspects of literacy that underpinned each test domain. This produced descriptions of increasing proficiency that reflected observations of student performance and a detailed analysis of the cognitive demands of PISA items. The second phase involved decisions about where to set cut-off points for levels and how to associate students with each level. This is both a technical and practical matter of interpreting what it means to be at a level, and has significant consequences for reporting national and international results.

Several principles were considered for developing and establishing a useful meaning for being at a level, and therefore for determining an approach to locating cut-off points between levels and associating students with them:

- A common understanding of the meaning of levels should be developed and promoted. First, it is important to understand that the literacy skills measured in PISA must be considered as continua: there are no natural breaking points to mark borderlines between stages along these continua. Dividing each of these continua into levels, though useful for communication about students' development, is essentially arbitrary. Like the definition of units on, for example, a scale of length, there is no fundamental difference between 1 metre and 1.5 metres – it is a matter of degree. It is useful, however, to define stages, or levels along the continua, because this enables communication about the proficiency of students in terms other than numbers. The approach adopted for PISA 2000 was that it would only be useful to regard students as having attained a particular level if this would allow certain expectations about what these students are capable of in general when they are said to be at that level. It was decided that this expectation would have to mean at a minimum that students at a particular level would be more likely to solve tasks at that level than to fail them. By implication, it must be expected that they would get at least half of the items correct on a test composed of items uniformly spread across that level, which is useful in helping to interpret the proficiency of students at different points across the proficiency range defined at each level.

  For example, students at the bottom of a level would complete at least 50 per cent of tasks correctly on a test set at the level, while students at the middle and top of each level would be expected to achieve a much higher success rate. At the top end of the bandwidth of a level would be the students who are masters of that level. These students would be likely to solve about 80 per cent of the tasks at that level. But, being at the top border of that level, they would also be at the bottom border of the next level up, where, according to the reasoning here, they should have a likelihood of at least 50 per cent of solving any tasks defined to be at that higher level.
- Further, the meaning of being at a level for a given scale should be more or less consistent for each level. In other words, to the extent possible within the substantively based definition and description of levels, cut-off points should create levels of more or less constant breadth. Some small variation may be appropriate, but in order for interpretation and definition of cut-off points and levels to be consistent, the levels have to be about equally broad. Clearly this would not apply to the highest and lowest proficiency levels, which are unbounded.

- A more or less consistent approach should be taken to defining levels for the different scales. Their breadth may not be exactly the same for the proficiency scales in different domains, but the same kind of interpretation should be possible for each scale that is developed.

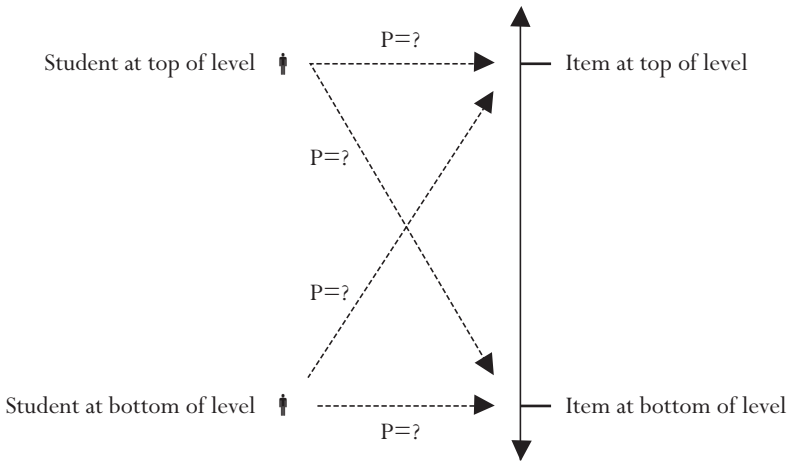
A way of implementing these principles was developed for PISA 2000 and it was used again for PISA 2003. This method links the two variables mentioned in the preceding dot-points, and a third related variable. The three variables can be expressed as follows:

- The expected success of a student at a particular level on a test containing items at that level (proposed to be set at a minimum that is near 50 per cent for the student at the bottom of the level, and higher for other students in the level);
- The width of the levels in that scale (determined largely by substantive considerations of the cognitive demands of items at the level and observations of student performance on the items); and
- The probability that a student in the middle of a level would correctly answer an item of average difficulty for that level (in fact, the probability that a student at any particular level would get an item at the same level correct), sometimes referred to as the "RP-value" for the scale (where "RP" indicates "response probability").

Figure 16.2 summarises the relationship among these three mathematically linked variables. It shows a vertical line representing a part of the scale being defined, one of the bounded levels on the scale, a student at both the top and the bottom of the level, and reference to an item at the top and an item at the bottom of the level. Dotted lines connecting the students and items are labelled P=? to indicate the probability associated with that student correctly responding to that item.

PISA 2000 implemented the following solution: start with the substantively determined range of abilities for each bounded level in each scale (the desired band breadth); then determine the highest possible RP-value that will be common across domains. That would give effect to the broad interpretation of the meaning of being at a level (an expectation of correctly responding to a minimum of 50 per cent of the items in a test at that level).

Figure 16.2 ■ **What it means to be at a level**



After doing this, the exact average percentage of correct answers on a test composed of items at a level could vary slightly among the different domains, but will always be at least 50 per cent at the bottom of the level.

The highest and lowest described levels are unbounded. For a certain high point on the scale and below a certain low point, the proficiency descriptions could, arguably, cease to be applicable. At the high end of the scale, this is not such a problem since extremely proficient students could reasonably be assumed to be capable of at least the achievements described for the highest level. At the other end of the scale, however, the same argument does not hold. A lower limit therefore needs to be determined for the lowest described level, below which no meaningful description of proficiency is possible.

As Levels 2, 3 and 4 (within a domain) will be equally broad, it was proposed that the floor of the lowest described level be placed at this breadth below the upper boundary of Level 1 (that is, the cut-off between levels 1 and 2). Student performance below this level is lower than that which PISA can reliably assess and, more importantly, describe.

REPORTING THE RESULTS FOR PISA MATHEMATICS

In this section, the way in which levels of mathematical literacy are defined, described and reported will be discussed. Levels of performance on the PISA mathematical literacy scale will be established and described, and they will be exemplified using a number of items from the PISA 2003 assessment.

Building an item map

The data from the PISA 2003 mathematics assessment were processed to generate a set of item difficulty measures for the 85 items included in the assessment. In fact, when the difficulty measures that were estimated for each of the partial credit steps of the polytomous items are also taken into account, a total of 93 item difficulty estimates were generated.

During the process of item development, experts undertook a qualitative analysis of each item, and developed descriptions of aspects of the cognitive demands of each item (and each individual item step in the case of partial credit items that were scored polytomously). This analysis included judgements about the aspects of the PISA mathematics framework that were relevant to each item. For example, each item was analysed to determine which of the overarching ideas was involved. Similarly, to establish the most apt competency cluster, the situation in which the stimulus and question were located was identified. As well as these broad categorisations, a short description was developed that attempted to capture the most important demands placed on students by each particular item, particularly the individual competencies that were called into play.

Following data analysis and the resultant generation of difficulty estimates for each of the 93 item steps, the items and item steps were associated with their difficulty estimates, with their framework classifications, and with their brief qualitative descriptions. Figure 16.3 shows a map of some of this information from a sample of items from the PISA 2003 test. Each row in Figure 16.3 represents an individual item or item step. The selected items and item steps have been ordered according to their difficulty, with the most difficult of these steps at the top, and the least difficult at the bottom. The difficulty estimate for each item and step is given, along with the associated classifications and descriptions.

Figure 16.3 ■ **A map for selected mathematics items**

Code	Item name	Item difficulty on PISA scale	Comments - item demands	Mathematics scale: Quantity	Mathematics scale: Space and shape	Mathematics scale: Change and relationships	Mathematics scale: Uncertainty	Competencies: Reproduction	Competencies: Connections	Competencies: Reflection	Context: Personal	Context: Educational / occupational	Context: Public	Context: Scientific
M124Q033	Walking Q3.3	723	Find a suitable strategy; multi-step problem solving; manipulation of expressions containing symbols; routine procedures; computations - multiply with decimals			■			■		■			
M179Q012	Robberies Q1.2	694	Interpret a graphical representation; construct a verbal explanation of a mathematical concept; mathematical argumentation skills based on use of data				■		■				■	
M266Q01	Carpenter Q1	687	Interpret and link text and diagrams representing a real-world situation; show insight in two-dimensional geometrical properties; extract information from geometrical representation; calculate		■				■			■		
M124Q032	Walking Q3.2	666	Find a suitable strategy; multi-step problem solving; manipulation of expressions containing symbols; routine procedures; partially correct computations			■			■		■			
M513Q01	Test Scores Q1	620	Look at a situation in a different way (statistics); link information in text and graph; establish a criterion and apply it; make use of simple statistical concepts; communicate argument in support of given proposition				■		■			■		
M124Q01	Walking Q1	611	Interpret and link picture, text and algebra; algebraic substitution; solve basic equation; single step; correct manipulation of expressions containing symbols			■		■			■			
M124Q031	Walking Q3.1	605	Find a suitable strategy; multi-step problem solving manipulation of expressions containing symbols; routine procedures; some computations - only first step carried out			■			■		■			
M413Q03	Exchange Rate Q3	586	Insight into quantitative relationships; strategy: how to tackle? (problem solving); communication of conclusion and reasoning	■						■			■	
M179Q011	Robberies Q1.1	577	Interpret a graphical representation; construct a partially correct verbal explanation of a mathematical concept; mathematical argumentation skills based on use of data				■		■				■	
M150Q03	Growing Up Q3	574	Interpret graph in respect to rate; reasoning; communicate explanation in support of given proposition			■			■					■
M520Q02	Skateboard Q2	570	Problem solving - choose a strategy; counting (combinatorics)	■				x			x			
M438Q02	Exports Q2	565	Interpret graph; identify and select relevant information; link separate data and carry out routine calculation				■		■				■	
M520Q03	Skateboard Q3	554	Explore possibilities to decide on which is best; interpret information; identify and select relevant information	■					■		■			
M150Q022	Growing Up Q2.2	525	Link text to graphical information; locate relevant data; write conclusion correctly			■		■						■
M555Q02	Number Cubes Q2	503	Spatial geometry; problem solving - devise a strategy; reasoning and insight - identify which are the pairs of opposite sides; apply given criteria in novel situation to evaluate scenarios		■				■		■			
M520Q012	Skateboard Q1.2	496	Interpret and link information in text and table; select and correctly process relevant information from a table; add all maximum values and all minimum values	■				■			■			
M150Q01	Growing Up Q1	477	Interpret graph and link to text; identify appropriate procedure carry out simple computation (subtraction)			■		■						■
M520Q011	Skateboard Q1.1	464	Interpret and link information in text and table; select and process relevant information from a table (only partially correctly)	■				■			■			
M413Q02	Exchange Rate Q2	439	Interpret simple quantitative model; apply it with a simple calculation (division)				■	■					■	
M438Q01	Exports Q1	427	Link representations (text and graphic); identify relevant information; read value directly from bar graph				■	■					■	
M547Q01	Staircase Q1	421	Interpret simple and familiar picture; simple calculation (division by two-digit number)		■			■				■		
M150Q021	Growing Up Q2.1	420	Link text to graphical information; locate relevant data; write a partially correct conclusion			■		■						■
M413Q01	Exchange Rate Q1	406	Interpret a simple quantitative model; apply it with a simple calculation involving multiplication	■				■					■	

When a map such as this is prepared using all available items, it becomes possible to look for factors that are associated with item difficulty. Many of those factors reflect variables that are central to constructs used in the mathematics framework's discussion of mathematical literacy. Indeed a very clear representation emerges of aspects of mathematical literacy that are associated with increasing item difficulty. Patterns emerge that make it possible to describe aspects of mathematical literacy that are consistently associated with various locations along the continuum shown by the map. For example, among the small sample of items in Figure 16.3, it can be seen that the easiest items are all from the reproduction competency cluster. This reflects the pattern observed with the full set of items. It is also seen from the full set of PISA items that those items characterised as belonging to the reflections cluster tend to be the most difficult. Items in the connections cluster tend to be of intermediate difficulty, though they span a large part of the proficiency spectrum that is analysed through the PISA assessment. In fact, the individual competencies defined in the mathematics framework play out quite differently at different levels of performance, in precisely the way that would be expected.

Near the bottom of the part of the continuum displayed here are items set in simple and relatively familiar contexts that require only the most limited amount of interpretation of the situation, and direct application of well-known mathematical knowledge in familiar situations. Typical activities are reading a value directly from a graph or table, performing a very simple and straightforward arithmetic calculation, ordering a small set of numbers correctly, counting familiar objects, using a simple currency exchange rate, identifying and listing simple combinatorial outcomes. For example, *Exchange Rate Q1* presents students with a simple rate for exchanging Singapore Dollars (SGD) into South African Rand (ZAR), namely 1 SGD = 4.2 ZAR. The question requires students to apply the rate to convert 3000 SGD into ZAR. The rate is presented in the form of a familiar equation, and the mathematical step required is direct and reasonably obvious.

Other examples, *Building Blocks Q1* and *Building Blocks Q2,* were presented in *The PISA 2003 Assessment Framework* (OECD, 2003). In those examples, students were presented with diagrams of a familiar three-dimensional shapes composed of small cubes, and asked to count (or calculate) the number of the small cubes used to make up the larger shapes.

Around the middle of the part of the continuum displayed are seen items that require substantially more interpretation, frequently of situations that are relatively unfamiliar or unpractised. They frequently demand the use of different representations of the situation, including more formal mathematical representations, and the thoughtful linking of those different representations in order to promote understanding and facilitate analysis. They often involve a chain of reasoning or a sequence of calculation steps, and can require expressing reasoning through a simple explanation. Typical activities are interpreting a set of related graphs; interpreting text, relating this to information in a table or graph, extracting the relevant information and performing some calculations; using scale conversions to calculate distances on a map; using spatial reasoning and geometric knowledge to perform distance, speed and time calculations. For example, the unit *Growing Up* presents students with a graph of the average height of young males and young females from the ages of 10 to 20 years. *Growing Up Q2* asks students to identify the period in their life when females are taller than males of the same age. Students have to interpret the graph to understand exactly what is being displayed; they have to relate the graphs for males and females to each other and determine how the specified period is shown then accurately read the relevant values from the horizontal scale. *Growing Up Q3* invites students to give a written explanation as to how the graph shows a slow-down in growth rate for girls after a particular age. To successfully answer this question, students must first understand how growth rate is displayed in such

a graph, must identify what is changing at the specified point in the graph in comparison to the period earlier than that, and must be able to articulate their explanation clearly in words.

Towards the top of the part of the scale displayed, can be seen items that typically involve a number of different elements and require even higher levels of interpretation. Situations are typically unfamiliar, hence requiring some degree of thoughtful reflection, and creativity. Questions usually demand some form of argumentation, often in the form of an explanation. Typical activities are interpreting complex and unfamiliar data; imposing a mathematical construction on a complex real-world situation; using mathematical modelling processes. At this part of the scale, items tend to have several elements that need to be linked by students, and their successful negotiation typically requires a strategic approach to several interrelated steps. For example, *Robberies Q1* presents students with a truncated bar graph showing the number of robberies per year in two specified years. A television reporter's statement interpreting the graph is given. Students are asked to consider whether or not the reporter's statement is a reasonable interpretation of the graph, and to give an explanation as to why. The graph itself is a little unusual, and requires some interpretation. The reporter's statement must be interpreted in relation to the graph. Then, some mathematical understanding and reasoning must be applied to determine a suitable meaning of the phrase 'reasonable interpretation' in this context. Finally, the conclusion must be articulated clearly in a written explanation. Fifteen-year-old students typically find such a sequence of thought and action quite challenging.

Another example illustrating items in this part of the mathematical literacy scale, *Heartbeat Q2*, was presented in the *PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills* (OECD, 2003). In that example, students were presented with mathematical formulations of the relationship between a person's recommended maximum heart rate, and their age, in the context of physical exercise. The question invited students to modify the formulation appropriately under a specified condition. They had to interpret the situation, the mathematical formulations, the changed condition, and construct a modified formulation that satisfied the specified condition. This complex set of linked tasks proved to be very challenging indeed.

Based on the patterns observed when the full item set is investigated in this way, growth along the PISA mathematical literacy scale can be characterised by referring to the ways in which mathematical competencies are associated with items located at different points along the scale.

The *PISA 2003 Assessment Framework* (OECD, 2003) summarises the following factors that underpin increasing levels of item difficulty and mathematical proficiency:

- The kind and degree of interpretation and reflection needed. This includes the nature of demands arising from the problem context; the extent to which the mathematical demands of the problem are apparent or to which students must impose their own mathematical construction on the problem; and the extent to which insight, complex reasoning and generalisation are required.
- The kind of representation skills that are necessary, ranging from problems where only one mode of representation is used, to problems where students have to switch between different modes of representation or to find appropriate modes of representation themselves.
- The kind and level of mathematical skill required, ranging from single-step problems requiring students to reproduce basic mathematical facts and perform simple computation processes through to multi-step

problems involving more advanced mathematical knowledge, complex decision making, information processing, and problem-solving and modelling skills.

- The kind and degree of mathematical argumentation that is required, ranging from problems where no arguing is necessary at all, through problems where students may apply well-known arguments, to problems where students have to create mathematical arguments or to understand other people's argumentation or judge the correctness of given arguments or proofs.

Levels of mathematical literacy

The approach to reporting used by the OECD following the PISA 2000 assessment of reading literacy was based on the definition of a number of bands or levels of reading literacy proficiency. Five levels were defined. Descriptions were developed to characterise typical student performance at each level. The levels were used to summarise the performance of students, to compare performances across subgroups of students, and to compare average performances among groups of students, in particular among the students from different participating countries. A similar approach has been used here to analyse and report PISA 2003 outcomes for mathematics.

For PISA mathematics, student scores have been transformed to the PISA scale, with a mean of 500 and a standard deviation of 100, and six levels of proficiency have been defined and described. The continuum of increasing mathematical literacy that is represented in Figure 16.2 has been divided into five bands, each of equal width, and two unbounded regions, one at each end of the continuum. The band definitions on the PISA scale are given in Table 16.1.

The information about the items in each band has been used to develop summary descriptions of the kinds of mathematical competencies associated with different levels of proficiency. These summary descriptions can then be used to encapsulate typical mathematical proficiency of students associated with each level. As a set, the descriptions encapsulate a representation of growth in mathematical literacy.

To develop the summary descriptions, growth in mathematical competence was first considered separately in relation to items from each of the four overarching ideas. Four sets of descriptions were developed. These are presented in following sections, in Figure 16.5 to Figure 16.8. The four sets of descriptions were combined to produce meta-descriptions of six levels of overall mathematical literacy, presented here in Figure 16.4.

Table 16.1 ■ **Mathematical literacy performance band definitions on the PISA scale**

Level	Score points on the PISA scale
6	Above 669
5	607 to 669
4	545 to 607
3	482 to 545
2	420 to 482
1	358 to 420

Figure 16.4 ■ **Summary descriptions for six levels of overall mathematical literacy**

	Overall mathematical literacy
6	At Level 6 students can conceptualise, generalise, and utilise information based on their investigations and modelling of complex problem situations. They can link different information sources and representations and flexibly translate among them. Students at this level are capable of advanced mathematical thinking and reasoning. These students can apply their insight and understandings along with a mastery of symbolic and formal mathematical operations and relationships to develop new approaches and strategies for attacking novel situations. Students at this level can formulate and precisely communicate their actions and reflections regarding their findings, interpretations, arguments, and the appropriateness of these to the original situations.
5	At Level 5 students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare, and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well-developed thinking and reasoning skills, appropriate linked representations, symbolic and formal characterisations, and insight pertaining to these situations. They can reflect on their actions and formulate and communicate their interpretations and reasoning.
4	At Level 4 students can work effectively with explicit models for complex concrete situations that may involve constraints or call for making assumptions. They can select and integrate different representations, including symbolic, linking them directly to aspects of real-world situations. Students at this level can utilise well-developed skills and reason flexibly, with some insight, in these contexts. They can construct and communicate explanations and arguments based on their interpretations, arguments and actions.
3	At Level 3 students can execute clearly described procedures, including those that require sequential decisions. They can select and apply simple problem-solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They can develop short communications reporting their interpretations, results and reasoning.
2	At Level 2 students can interpret and recognise situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions. They are capable of direct reasoning and making literal interpretations of the results.
1	At Level 1 students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are obvious and follow immediately from the given stimuli.

A clear progression through these levels is apparent in the way in which the individual mathematical competencies specified in the PISA mathematics framework play out as literacy levels increase.

For example, the competency thinking and reasoning is observed to follow a progression through six stages:

1. Follow direct instructions and take obvious actions.
2. Use direct reasoning and literal interpretations.
3. Make sequential decisions, interpret and reason from different information sources.
4. Employ flexible reasoning and some insight.
5. Use well developed thinking and reasoning skills.
6. Use advanced mathematical thinking and reasoning.

The communication competency follows this progression:

1. Follow explicit instructions.
2. Extract information and make literal interpretations.
3. Produce short communications supporting interpretations.
4. Construct and communicate explanations and argument.
5. Formulate and communicate interpretations and reasoning.
6. Formulate precise communications.

For modelling, the following development is observed as literacy levels increase:

1. Apply simple given models.
2. Recognise, apply and interpret basic given models.
3. Make use of different representational models.
4. Work with explicit models, and related constraints and assumptions.
5. Develop and work with complex models; reflect on modelling processes and outcomes.
6. Conceptualise and work with models of complex mathematical processes and relationships; reflect on, generalise and explain modelling outcomes.

For problem posing and problem solving, the following development is observed as literacy levels increase:

1. Handle direct and explicit problems.
2. Use direct inference.
3. Use simple problem-solving strategies.
4. Work with constraints and assumptions.
5. Select, compare and evaluate appropriate problem-solving strategies.
6. Investigate and model with complex problem situations.

In the case of the competency representation, the following development is observed as literacy levels increase:

1. Handle familiar and direct information.
2. Extract information from single representations.
3. Interpret and use different representations.
4. Select and integrate different representations and link them to real world situations.
5. Make strategic use of appropriately linked representations.
6. Link different information and representations and translate flexibly among them.

Likewise, for using symbolic, formal and technical language and operations, these stages of development are observed:

1. Apply routine procedures.
2. Employ basic algorithms, formulae, procedures and conventions.
3. Work with symbolic representations.
4. Use symbolic and formal characterisations.
5. Mastery of symbolic and formal mathematical operations and relationships.

The following four figures (Figure 16.5, Figure 16.6, Figure 16.7, and Figure 16.8) show proficiency descriptions for each of the six levels for each of the overarching ideas of the mathematics framework. These descriptions comprise a summary description for the mathematical content area, and a set of more detailed illustrative statements that relate to competencies required by the items located at each level.

Figure 16.5 ■ **Summary descriptions of six levels of proficiency in the *Quantity* area**

	Summary description	Illustrative competencies
6	Conceptualise and work with models of complex mathematical processes and relationships; work with formal and symbolic expressions; use advanced reasoning skills to devise strategies for solving problems and to link multiple contexts; use sequential calculation processes; formulate conclusions, arguments and precise explanations.	– Conceptualise complex mathematical processes such as exponential growth, weighted average, as well as number properties and numeric relationships – Interpret and understand complex information, and link multiple complex information sources – Use advanced reasoning concerning proportions, geometric representations of quantities, combinatorics and integer number relationships – Interpret and understand formal pure mathematical expressions of relationships among numbers, including in a scientific context – Perform sequential calculations in a complex and unfamiliar context, including working with large numbers – Formulate conclusions, arguments and precise explanations – Devise a strategy (develop heuristics) for working with complex mathematical processes
5	Work effectively with models of more complex situations to solve problems; use well-developed reasoning skills, insight and interpretation with different representations; carry out sequential processes; communicate reasoning and argument.	– Interpret complex information about real-world situations (including graphs, drawings and complex tables) – Link different information sources (such as graphs, tabular data and related text) – Extract relevant data from a description of a complex situation and perform calculations – Use problem-solving skills (*e.g.* interpretation, devising a strategy, reasoning; systematic counting) in real-world contexts that involve substantial mathematisation – Communicate reasoning and argument – Make an estimation using daily life knowledge – Calculate relative and/or absolute change
4	Work effectively with simple models of complex situations; use reasoning skills in a variety of contexts, interpret different representations of the same situation; analyse and apply quantitative relationships; use a variety of calculation skills to solve problems.	– Accurately apply a given numeric algorithm involving a number of steps – Interpret complex text descriptions of a sequential process – Relate text-based information to a graphic representation – Perform calculations involving proportional reasoning, divisibility or percentages in simple models of complex situations – Perform systematic listing and counting of combinatorial outcomes – Identify and use information from multiple sources – Analyse and apply a simple system – Interpret complex text to produce a simple mathematical model
3	Use simple problem-solving strategies including reasoning in familiar contexts; interpret tables to locate information; carry out explicitly described calculations including sequential processes.	– Interpret a text description of a sequential calculation process, and correctly implement the process – Use basic problem-solving processes (devise a simple strategy, look for relationships, understand and work with given constraints, use trial and error, simple reasoning) – Perform calculations including working with large numbers, calculations with speed and time, conversion of units (*e.g.* from annual rate to daily rate) – Interpret tabular information, locate relevant data from a table – Conceptualise relationships involving circular motion and time – Interpret text and diagram describing a simple pattern
2	Interpret simple tables to identify and extract relevant information; carry out basic arithmetic calculations; interpret and work with simple quantitative relationships.	– Interpret a simple quantitative model (*e.g.* a proportional relationship) and apply it using basic arithmetic calculations – Interpret simple tabular data, link textual information to related tabular data – Identify the simple calculation required to solve a straight-forward problem – Perform simple calculations involving the basic arithmetic operations, as well as ordering numbers
1	Solve problems of the most basic type in which all relevant information is explicitly presented, the situation is straightforward and very limited in scope, the required computational activity is obvious and the mathematical task is basic, such as a simple arithmetic operation.	– Interpret a simple, explicit mathematical relationship, and apply it directly using a calculation – Read and interpret a simple table of numbers, total the columns and compare the results

Figure 16.6 ■ **Summary descriptions of six levels of proficiency in the *Space and shape* area**

	Summary description	Illustrative competencies
6	Solve complex problems involving multiple representations and often involving sequential calculation processes; identify and extract relevant information and link different but related information; use reasoning, significant insight and reflection; generalise results and findings, communicate solutions and provide explanations and argumentation.	– Interpret complex textual descriptions and relate these to other (often multiple) representations – Use reasoning involving proportions in non-familiar and complex situations – Show significant insight to conceptualise complex geometric situations or to interpret complex and unfamiliar representations – Identify and combine multiple pieces of information to solve problems – Devise a strategy to connect a geometrical context with known mathematical procedures and routines – Carry out a complex sequence of calculations (*e.g.* volume calculations or other routine procedures in an applied context) accurately and completely – Provide written explanations and argument based on reflection, insight and generalisation of understandings
5	Solve problems that require appropriate assumptions to be made, or that involve working with assumptions provided; use well-developed spatial reasoning, argument and insight to identify relevant information and to interpret and link different representations; work strategically and carry out multiple and sequential processes.	– Use spatial/geometrical reasoning, argument, reflection and insight into two- and three-dimensional objects, both familiar and unfamiliar – Make assumptions or work with assumptions to simplify and solve a geometrical problem in a real-world setting (*e.g.* involving estimation of quantities in a real-world situation) and communicate explanations – Interpret multiple representations of geometric phenomena – Use geometric constructions – Conceptualise and devise multi-step strategy to solve geometrical problems – Use well-known geometrical algorithms but in unfamiliar situations, such as Pythagoras's theorem; and calculations involving perimeter, area and volume
4	Solve problems that involve visual and spatial reasoning and argumentation in unfamiliar contexts; link and integrate different representations; carry out sequential processes; apply well-developed skills in spatial visualisation and interpretation.	– Interpret complex text to solve geometric problems – Interpret sequential instructions; follow a sequence of steps – Interpretation using spatial insight into non-standard geometric situations – Use a two-dimensional model to work with three-dimensional representations of unfamiliar geometric situation – Link and integrate two different visual representations of a geometric situation – Develop and implement a strategy involving calculation in geometric situations – Reasoning and argument about numeric relationships in a geometric context – Perform simple calculations *(e.g.* multiply multi-digit decimal number by an integer, numeric conversions using proportion and scale, calculate areas of familiar shapes)
3	Solve problems that involve elementary visual and spatial reasoning in familiar contexts; link different representations of familiar objects; use elementary problem-solving skills (devising simple strategies); apply simple algorithms.	– Interpret textual descriptions of unfamiliar geometric situations – Use basic problem-solving skills, such as devising a simple strategy – Use visual perception and elementary spatial reasoning skills in a familiar situation – Work with a given familiar mathematical model – Perform simple calculations such as scale conversions (using multiplication, basic proportional reasoning) – Apply routine algorithms to solve geometric problems (*e.g.* calculate lengths within familiar shapes)
2	Solve problems involving a single mathematical representation where the mathematical content is direct and clearly presented; use basic mathematical thinking and conventions in familiar contexts.	– Recognise simple geometric patterns – Use basic technical terms and definitions and apply basic geometric concepts (*e.g.* symmetry) – Apply a mathematical interpretation of a common-language relational term (*e.g.* "bigger") in a geometric context – Create and use a mental image of an object, both two- and three-dimensional – Understand a visual two-dimensional representation of a familiar real-world situation – Apply simple calculations (*e.g.* subtraction, division by a two-digit number) to solve problems in a geometric setting
1	Solve simple problems in a familiar context using familiar pictures or drawings of geometric objects and applying counting or basic calculation skills.	– Use a given two-dimensional representation to count or calculate elements of a simple three-dimensional object

Figure 16.7 ■ **Summary descriptions of six levels of proficiency in the *Change and relationships* area**

	Summary description	Illustrative competencies
6	Use significant insight, abstract reasoning and argumentation skills and technical knowledge and conventions to solve problems and to generalise mathematical solutions to complex real-world problems.	– Interpret complex mathematical information in the context of an unfamiliar real-world situation – Interpret periodic functions in a real-world setting, perform related calculations in the presence of constraints – Interpret complex information hidden in the context of an unfamiliar real-world situation – Interpret complex text and use abstract reasoning (based on insight into relationships) to solve problems – Insightful use of algebra or graphs to solve problems; ability to manipulate algebraic expressions to match a real-world situation – Undertake problem solving based on complex proportional reasoning – Implement multi-step problem-solving strategies involving the use of formula and calculations – Devise a strategy and solve the problem by using algebra or trial-and-error – Identify a formula which describes a complex real-world situation, generalise exploratory findings to create a summarising formula – Generalise exploratory findings in order to do some calculations – Apply deep geometrical insight to work with and generalise complex patterns – Conceptualise complex percentage calculation – Coherently communicate logical reasoning and arguments
5	Solve problems by making advanced use of algebraic and other formal mathematical expressions and models. Link formal mathematical representations to complex real-world situations. Use complex and multi-step problem-solving skills, reflect on and communicate reasoning and arguments.	– Interpret complex formulae in a scientific context – Interpret periodic functions in a real-world setting, and perform related calculations – Use advanced problem-solving strategies: Interpret and link complex information; Interpret and apply constraints; – Identify and carry out a suitable strategy – Reflect on the relationship between an algebraic formula and its underlying data – Use complex proportional reasoning (*e.g.* related to rates) – Analyse and apply a given formula in a real-life situation – Communicate reasoning and argument
4	Understand and work with multiple representations, including explicitly mathematical models of real-world situations to solve practical problems. Employ considerable flexibility in interpretation and reasoning, including in unfamiliar contexts, and communicate the resulting explanations and arguments.	– Interpret complex graphs, and read one or multiple values from graphs – Interpret complex and unfamiliar graphical representations of real-world situations – Use multiple representations to solve a practical problem – Relate text-based information to a graphic representation and communicate explanations – Analyse a formula describing a real-world situation – Analyse three-dimensional geometric situations involving volume and related functions – Analyse a given mathematical model involving a complex formula – Interpret and apply word formulae, and manipulate and use linear formulae that represent real-world relationships – Carry out a sequence of calculations involving percentage, proportion, addition or division
3	Solve problems that involve working with multiple related representations (text, graph, table, formulae), including some interpretation, reasoning in familiar contexts, and communication of argument.	– Interpret unfamiliar graphical representations of real-world situations – Identify relevant criteria in a text – Interpret text in which a simple algorithm is hidden and apply that algorithm – Interpret text and devise a simple strategy – Link and connect multiple related representations (*e.g.* two related graphs, text and a table, a formula and a graph) – Use reasoning involving proportions in various familiar contexts and communicate reasons and argument – Apply a text, given criterion or situation, to a graph – Use a range of simple calculation procedures to solve problems, including ordering data, time difference calculations, linear interpolation
2	Work with simple algorithms, formulae and procedures to solve problems; link text with a single representation (graph, table, simple formula); use interpretation and reasoning skills at an elementary level.	– Interpret a simple text and link it correctly to graphical elements – Interpret a simple text that describes a simple algorithm and apply that algorithm – Interpret a simple text and use proportional reasoning or a calculation – Interpret a simple pattern – Interpret and use reasoning in a practical context involving a simple and familiar application of motion, speed and time relationships – Locate relevant information in graph, and read values directly from the graph – Correctly substitute numbers to apply a simple numeric algorithm or simple algebraic formula
1	Locate relevant information in a simple table or graph; follow direct and simple instructions to read information directly from a simple table or graph in a standard or familiar form; perform simple calculations involving relationships between two familiar variables.	– Make a simple connection of text to a specific feature of a simple graph and read off a value from the graph – Locate and read a specified value in a simple table – Perform simple calculations involving relationships between two familiar variables

© OECD 2005

Figure 16.8 ■ **Summary descriptions of six level of proficiency in the *Uncertainty* area**

	Summary Description	Illustrative competencies
6	Use high-level thinking and reasoning skills in statistical or probabilistic contexts to create mathematical representations of real-world situations; use insight and reflection to solve problems and to formulate and communicate arguments and explanations.	– Interpret and reflect on real world situations using probability knowledge and carry out resulting calculations using proportional reasoning, large numbers and rounding – Show insight into probability in a practical context – Use interpretation, logical reasoning and insight at a high level in an unfamiliar probabilistic situation – Use rigorous argumentation based on insightful interpretation of data – Employ complex reasoning using statistical concepts – Show understanding of basic ideas of sampling and carry out calculations with weighted averages, or using insightful systematic counting strategies – Communicate complex arguments and explanations
5	Apply probabilistic and statistical knowledge in problem situations that are somewhat structured and where the mathematical representation is partially apparent. Use reasoning and insight to interpret and analyse given information, to develop appropriate models and to perform sequential calculation processes; communicate reasons and arguments.	– Interpret and reflect on the outcomes of an unfamiliar probabilistic experiment – Interpret text using technical language and translate to an appropriate probability calculation – Identify and extract relevant information, and interpret and link information from multiple sources (e.g. from text, multiple tables, graphs) – Use reflection and insight into standard probabilistic situations – Apply probability concepts to analyse a non-familiar phenomenon or situation – Use proportional reasoning and reasoning with statistical concepts – Use multi-step reasoning based on data – Carry out complex modelling involving the application of probability knowledge and statistical concepts (*e.g.* randomness, sample, independence) – Use calculations including addition, proportions, multiplication of large numbers, rounding, to solve problems in non-trivial statistical contexts – Carry out a sequence of related calculations – Carry out and communicate probabilistic reasoning and argument
4	Use basic statistical and probabilistic concepts combined with numerical reasoning in less familiar contexts to solve simple problems; carry out multi-step or sequential calculation processes; use and communicate argumentation based on interpretation of data.	– Interpret text, including in an unfamiliar (scientific) but straight-forward context – Show insight into aspects of data from tables and graphs – Translate text description into appropriate probability calculation – Identify and select data from various statistical graphs and carry out basic calculation – Show understanding of basic statistical concepts and definitions (probability, expected value, randomness, average) – Use knowledge of basic probability to solve problems – Construct a basic mathematical explanation of a verbal real-world quantitative concept (*e.g.* "huge increase") – Use mathematical argumentation based on data – Use numerical reasoning – Carry out multi-step calculations involving the basic arithmetic operations, and working with percentage – Draw information from a table and communicate a simple argument based on that information
3	Interpret statistical information and data, and link different information sources; basic reasoning with simple probability concepts, symbols and conventions and communication of reasoning.	– Interpret tabular information – Interpret and read from non-standard graphs – Use reasoning to identify probability outcomes in the context of a complex but well-defined and familiar probability experiment – Insight into aspects of data presentation (*e.g.* number sense, link related information from two different tables), link data to suitable chart type – Communicate common-sense reasoning
2	Locate statistical information presented in familiar graphical form; understand basic statistical concepts and conventions.	– Identify relevant information in a simple and familiar graph – Link text to a related graph, in a common and familiar form – Understand and explain simple statistical calculations (the average) – Read values directly from a familiar data display, such as a bar graph
1	Understand and use basic probabilistic ideas in familiar experimental contexts.	– Understand basic probability concepts in the context of a simple and familiar experiment (*e.g.* involving dice or coins) – Systematic listing and counting of combinatorial outcomes in a limited and well-defined game situation

Interpreting the mathematical literacy levels

The proficiency levels defined and described in the preceding section require one more set of technical decisions before they can be used to summarise and report the performance of particular students. The scale of PISA mathematical literacy is a continuous scale. The use of performance bands, or levels of proficiency, involves an essentially arbitrary division of that continuous scale into discrete parts. The number of divisions, and the location of the cut-points that mark the boundaries of the divisions, are two matters that must be determined. For PISA mathematics, the scale has been divided into seven regions, including 5 bounded regions labelled levels 1 to 5, an unbounded region below level 1, and an unbounded upper region (labelled level 6). The cutpoints that mark the boundaries between these regions were given in Table 16.1.

The creation of these performance bands leads to a situation where a range of values on the continuous scale is grouped together into each single band. Given that range of performances within each level, how are individual students assigned to the levels and meaning can be ascribed to being at a level? In the context of the OECD reporting of PISA 2000 results, a commonsense interpretation of the meaning of being at a level was developed and adopted. That is, students are assigned to the highest level for which they would be expected to correctly answer the majority of assessment items. Imagine a test composed of items spread uniformly across a level: a student near the bottom of the level will be expected to correctly answer at least half of the test questions from that level. Students at progressively higher points in that level would be expected to correctly answer progressively more of the questions in that level. It should be remembered that the relationship between students and items is probabilistic – it is possible to estimate the probability that a student at a particular location on the scale will get an item at a particular location on the scale correct. Students assigned to a particular level will be expected to successfully complete some items from the next higher level, and it is only when that expectation reaches the threshold of at least half of the items in the next higher level that the student would be placed in the next higher level. Mathematically, the probability level used to assign students to the scale to achieve this commonsense interpretation of being at a level is 0.62. Students are placed on the scale at the point where they have a 62 per cent chance of correctly answering test questions located at the same point.

The same meaning has been applied in the reporting of PISA 2003 results. Such an approach makes it possible to summarise aspects of student proficiency by describing the things related to PISA mathematical literacy that students can be expected to do at different locations on the scale.

REPORTING THE RESULTS FOR PISA PROBLEM SOLVING

Cross-curricular problem-solving competencies were included in the PISA 2003 assessment as a minor assessment domain. Details of the problem-solving domain are provided in the *PISA 2003 Assessment Framework* (OECD, 2003). The amount of data available from the limited number of test items used in the assessment was much smaller than was the case for mathematics. Therefore it was not possible to develop fully detailed proficiency descriptions for a full range of levels. However, a process similar to that described for PISA mathematics in the previous section was also applied to the development of a described scale for problem-solving proficiency. The outcomes of that development and of the problem-solving assessment are described in detail in *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003* (OECD, 2004b).

Problem-solving proficiency was described in three broad levels:

Level 3: Reflective, communicative problem solvers

Students proficient at Level 3 score above 592 points on the PISA problem-solving scale and typically do not only analyse a situation and make decisions, but also think about the underlying relationships in a problem and relate these to the solution. Students at Level 3 approach problems systematically, construct their own representations to help them solve it and verify that their solution satisfies all requirements of the problem. These students communicate their solutions to others using accurate written statements and other representations.

Students at Level 3 tend to consider and deal with a large number of conditions, such as monitoring variables, accounting for temporal restrictions, and other constraints. Problems at this level are demanding and require students to regulate their work. Students at the top of Level 3 can cope with multiple interrelated conditions that require students to work back and forth between their solution and the conditions laid out in the problem. Students at this level organise and monitor their thinking while working out their solution. Level 3 problems are often multi-faceted and require students to manage all interactions simultaneously and develop a unique solution, and students at Level 3 are able to address such problems successfully and communicate their solutions clearly.

Students at Level 3 are also expected to be able to successfully complete tasks located at lower levels of the PISA problem-solving scale.

Level 2: Reasoning, decision-making problem solvers

Students proficient at Level 2 score from 500 to 592 points on the problem-solving scale and use reasoning and analytic processes and solve problems requiring decision making skills. These students can apply various types of reasoning (inductive and deductive reasoning, reasoning about causes and effects, or reasoning with many combinations, which involves systematically comparing all possible variations in well-described situations) to analyse situations and to solve problems that require them to make a decision among well-defined alternatives. To analyse a system or make decisions, students at Level 2 combine and synthesise information from a variety of sources. They are able to combine various forms of representations (*e.g.* a formalised language, numerical information, and graphical information), handle unfamiliar representations (*e.g.* statements in a programming language, or flow diagrams related to a mechanical or structural arrangement of components) and draw inferences based on two or more sources of information.

Students at Level 2 are also expected to be able to successfully complete tasks located at Level 1 of the PISA problem-solving scale.

Level 1: Basic problem solvers

Students proficient at Level 1 score from 405 to 499 points on the problem-solving scale and typically solve problems where they have to deal with only a single data source containing discrete, well-defined information. They understand the nature of a problem and consistently locate and retrieve information related to the major features of the problem. Students at Level 1 are able to transform the information in the problem to present the problem differently, *e.g.* take information from a table to create a drawing or graph. Also, students can apply information to check a limited number of well-defined conditions within the problem. However, students at Level 1 do not typically deal successfully with multi-faceted problems involving more than one data source or requiring them to reason with the information provided.

Below Level 1: Weak or emergent problem solvers

The PISA problem-solving assessment was not designed to assess elementary problem-solving processes. As such, the assessment materials did not contain sufficient tasks to describe fully performances that fall below Level 1. Students with performances below Level 1 have scores of less than 405 points on the problem-solving scale and consistently fail to understand even the easiest items in the assessment or to apply the necessary processes to characterise important features or represent the problems.

At most, they can deal with straightforward problems with carefully structured tasks that require students to give responses based on facts or to make observations with few or no inferences. Students below Level 1 have significant difficulties in making decisions, analysing or evaluating systems, and trouble-shooting situations.

REPORTING THE RESULTS FOR PISA READING AND SCIENCE

The scales used for the reporting of reading and science in PISA 2003 are identical to those used in PISA 2000. Details on the reading and science scales can be found in the *PISA 2000 Technical Report* (OECD, 2002).

Note

1 While strictly speaking the scales based on aspects of reading are subscales of the combined reading literacy scale, for simplicity they are mostly referred to as 'scales' rather than 'subscales' in this report.

17

Scaling Procedures and Construct Validation of Context Questionnaire Data

OVERVIEW

The PISA 2003 context questionnaires included numerous items on student characteristics, student family background, student perceptions, school characteristics and school principals' perceptions. Though some of these questions can be analysed as single items (for example, gender), most questions were designed to measure latent constructs that cannot be observed directly. Here, transformations or scaling procedures are needed to construct meaningful indices.

This chapter describes how student and school questionnaire indices were constructed and validated. In PISA 2003, two types of indices can be distinguished:

- *Simple indices:* These indices were constructed through the arithmetical transformation or recoding of one or more items. Item responses were used to calculate meaningful variables.
- *Scale indices:* Variable construction through the scaling of items. Typically, scale scores for these indices are estimates of latent traits derived through IRT (item response theory) scaling of dichotomous or Likert-type items.

This chapter outlines how simple indices were constructed; describes the methodology used for construct validation, scaling and scale description; details the construction and validation of scaled indices; and illustrates the computation of the index on economic, social and cultural status (ESCS). Some of the indices had already been used in PISA 2000 and the scaling methodology is similar to that used in the PISA 2000 (OECD, 2002). Most indices, however, were based on the elaboration of a questionnaire framework and are related to mathematics as the major domain of PISA 2000 (see Chapter 3).

SIMPLE QUESTIONNAIRE INDICES

Student indices

Student age

The age of a student (AGE) was calculated as the difference between the year and month of the testing and the year and month of a student's birth. Data on students' age were obtained both from the questionnaire and the student tracking forms.

Relative grade of student

Data on the student's grade are obtained both from the student questionnaire and from the student tracking forms. In order to adjust for between-country variation the index of the students' relative grade (GRADE) indicates whether students are at the modal grade in a country (value of 0), or whether they are below or above the modal grade (+x grades, –x grades).

Study programme indices

PISA 2003 collected data on study programmes available to 15-year-old students in each country. At the individual level the study programme was identified both through the student tracking form and the student questionnaire. All study programmes were classified using the International Standard Classification of Education (ISCED) (OECD, 1999). The following indices are derived from the data on study programmes: programme level (ISCEDL) indicating whether students are on the lower or upper secondary level (ISCED 3 or ISCED 2); programme designation (ISCEDD) indicating the designation of the study programme

(A = general programmes designed to give access to the next programme level, B = programmes designed to give access to vocational studies at the next programme level, C = programmes designed to give direct access to the labour market, M = modular programmes that combine any or all of these characteristics); and programme orientation (ISCEDO) indicating whether the programme's curricular content is general, pre-vocational or vocational.

Family structure

Student reports on who is living with them at home were recoded into an index of family structure (FAMSTRUC) with four categories: (1) is a single parent family (students living with only one of the following: mother, female guardian, father, male guardian); (2) is a nuclear family (students living with a father and a mother); (3) is a mixed family (a father and a guardian, a mother and a guardian, or two guardians); and (4) groups the other responses, except the non-responses which were coded as missing or not applicable.

Highest occupational status of parents

Occupational data for both the student's father and student's mother were obtained by asking open-ended questions. The responses were coded to four-digit ISCO codes (ILO, 1990) and then mapped to the international socio-economic index of occupational status (ISEI) (Ganzeboom *et al.,* 1992). Three indices were obtained from these scores: father's occupational status (BFMJ); mother's occupational status (BMMJ); and the highest occupational status of parents (HISEI) which corresponds to the higher ISEI score of either parent or to the only available parent's ISEI score. For all three indices, higher ISEI scores indicate higher levels of occupational status.

Educational level of parents

Parental education is a second family background variable that is often used in the analysis of educational outcomes. Theoretically, it has been argued that parental education is a more relevant influence on student's outcomes than is parental occupation. Like occupation, the collection of internationally comparable data on parental education poses significant challenges, and less work has been done on internationally comparable measures of educational outcomes than has been done on occupational status. The core difficulties with parental education relate to international comparability (education systems differ widely between countries and within countries over time) and response validity (students are often unable to accurately report their parents' levels of education).

In PISA, parental education is classified using ISCED (OECD, 1999). Indices on parental education are constructed by recoding educational qualifications into the following categories: (0) None; (1) ISCED 1 (primary education); (2) ISCED 2 (lower secondary); (3) ISCED Level 3B or 3C (vocational/pre-vocational upper secondary); (4) ISCED 3A (upper secondary) and/or ISCED 4 (non-tertiary post-secondary); (5) ISCED 5B (vocational tertiary); and (6) ISCED 5A, 6 (theoretically oriented tertiary and post-graduate). Indices with these categories were provided for the students' mother (MISCED) and the students' father (FISCED) of the student. In addition, the index on the highest educational level of parents (HISCED) corresponds to the higher ISCED level of either parent.

The index scores for highest educational level of parents were also recoded into estimated years of schooling (PARED). A mapping of ISCED levels to years of schooling is provided in Appendix 16.

Immigration background

As in PISA 2000, information on the country of birth of the students and their parents was collected. The index on immigrant background (IMMIG) was already used in PISA 2000 and has the following categories: (1) native students (those students born in the country of assessment or who had at least one parent born in the country);[1] (2) first-generation students (those born in the country of assessment but whose parents were both born in another country; and (3) non-native students (those students born outside the country of assessment and whose parents were also born in another country). Students with missing responses for either the student or for both parents, or for all three questions were given missing values.

Language spoken at home

Students were asked if the language spoken at home most of the time was the language of assessment, another official national language, another national dialect or language, or another language (some countries collected more detailed information on language use, which is included in the database with international language codes). In order to derive an index of language spoken at home (LANG), responses were grouped into two categories: (1) language spoken at home most of the time is different from the language of assessment, from other official national languages and from other national dialects or languages; and (0) the language spoken at home most of the time is the language of assessment, is another official national language, or other national dialect or language.

Expected educational level

In PISA 2003 students were asked about their educational aspirations. Educational levels were classified according to ISCED (OECD, 1999). The index of the expected educational level (SISCED) has the following categories: (1) ISCED 1 (students not expecting to finish their current ISCED 2 programme); (2) ISCED 2 (lower secondary); (3) ISCED Level 3B or 3C (vocational/prevocational upper secondary); (4) ISCED 3A (upper secondary) or ISCED 4 (non-tertiary post-secondary); (5) ISCED 5B (vocational tertiary); and (6) ISCED 5A, 6 (theoretically oriented tertiary and post-graduate).

Expected occupational status

As part of the optional questionnaire on Educational Career, students in 24 countries were asked to report their expected occupation and a description of this job. The responses were coded to four-digit ISCO codes (ILO, 1990) and then mapped to the ISEI index (Ganzeboom *et al.,* 1992). Recoding of ISCO codes into ISEI index results in scores for the students' expected occupational status (BSMJ), where higher scores of ISEI indicate higher level of expected occupational status.

Mathematics homework

As in PISA 2000, students were asked about the amount of homework that they did. For PISA 2003, they were asked about their mathematics and their overall homework in hours. The ratio of time spent on mathematics homework and the overall time spent on homework provides an index of relative time spent on mathematics homework (RMHMWK).

Instructional time

Students were asked to provide information on the average length of a class period and their instructional time in mathematics in class periods. Two major problems had to be dealt with when collecting data on instructional time:

- Students were asked about the average length of a class period, because in numerous countries there is variation both across and within schools. However, individual estimates did not provide a sufficient degree of reliability. Therefore, after applying plausibility checks and discarding student values that seemed unreasonable, the median of reported class length was calculated for each study programme within schools. These aggregated numbers were then used to compute the instructional time for individual students.
- In some countries the amount of instructional time in mathematics varies across the year. Therefore, some students may have attended mathematics lessons at the time of the testing but not at other times of the year, or may have instruction in this subject earlier in the school year but not at the time of testing. Therefore, the information captured by this question refers only to the current instruction in mathematics received by each tested student.

Three indices on instructional time were derived: the index of minutes of overall school instruction (TMINS) is calculated by multiplying the median length of a class period (at the level of study programmes within schools) by the number of class periods with instruction in all subjects (including mathematics) as reported by the student; the index of minutes of mathematics instruction (MMINS) is calculated by multiplying the median length of a class period (at the level of study programmes within schools) by the number of class periods receiving mathematics instruction; and the index of the relative instructional time on mathematics (PCMATH) is calculated by dividing the instructional time in minutes on mathematics (MMINS) by the overall instructional time in minutes (TMINS).

School indices

School size

The PISA 2000 and PISA 2003 index of school size (SCHLSIZE) contains the total enrolment at school based on the enrolment data provided by the school principal, summing the number of girls and boys at a school.

Proportion of girls enrolled at school

The PISA 2000 and PISA 2003 index of the proportion of girls enrolled at school (PCGIRLS) is based on the enrolment data provided by the school principal, dividing the number of girls by the total of girls and boys at a school.

School type

Schools are classified as either public or private according to whether a private entity or a public agency has the ultimate power to make decisions concerning its affairs. This PISA 2000/2003 index of school type (SCHLTYPE) has three categories: (1) public schools controlled and managed by a public education authority or agency; (2) government-dependent private schools controlled by a non-government organisation or with a governing board not selected by a government agency, but which receive more than 50 per cent of their core funding from government agencies; and (3) government-independent private schools controlled by a non-government organisation or with a governing board not selected by a government agency and which receive less than 50 per cent of their core funding from government agencies.[2]

Availability of computers

School principals were asked to report the number of computers available at school. The index of availability of computers (RATCOMP) is obtained by dividing the number of computers at school by the number of students at school.

In addition, the index of proportion of computers connected to Web (COMPWEB) and the index of proportion of computers connected to a Local Area Network (COMPLAN) were computed. The former is the number of computers connected to the Web divided by the total number of computers and the latter is the number of computers connected to a local network divided by the total number of computers.

Quantity of teaching staff at school

School principals were asked to report the number of full-time and part-time teachers at school. Teachers in general and mathematics teachers were reported separately. For all of the following indices the number of part-time teachers contributed 0.5 and the number of full-time teachers 1.0 to the estimated numbers of teachers at school.

- The index of student/teacher ratio (STRATIO) was obtained by dividing the number of enrolled students (index SCHLSIZE) by the total number of teachers.
- The index of proportion of fully certified teachers (PROPCERT) was computed by dividing the number of fully certified teachers by the total number of teachers.
- The index of proportion of teachers with an ISCED 5A qualification in pedagogy (PROPQPED) was calculated by dividing the number of teachers with this qualification by the total number of teachers.
- The index of student/mathematics teacher ratio (SMRATIO) was obtained by dividing the number of enrolled students (SCHLSIZE) by the total number of mathematics teachers.
- The index of proportion of mathematics teachers (PROPMATH) was computed by dividing the number of mathematics teachers by the total number of teachers.
- The index of proportion of mathematics teachers with an ISCED 5A qualification and a major in mathematics (PROPMA5A) was calculated by dividing the number of the mathematics teachers with this qualification by the total number of mathematics teachers.

School selectivity

School principals were asked about admittance policies at their school. Among these policies, principals were asked how much consideration was given to the following factors when students are admitted to the school, based on a scale with the categories "not considered", "considered", "high priority" and "prerequisite": students' academic record (including placement tests) and the recommendation of feeder schools. An index of school selectivity (SELECT) was computed by assigning schools to four different categories: (1) schools where none of these factors is considered for student admittance; (2) schools considering at least one of these factors; (3) schools giving high priority to at least one of these factors; and (4) schools where at least one of these factors is a pre-requisite for student admittance.

Use of assessments

School principals were asked to rate the frequency of the following assessments for 15-year-old students at school: i) standardised tests; ii) teacher-developed tests; iii) teachers' judgemental ratings; iv) student portfolios; and v) student assignments/projects/homework. All five items were recoded into numerical values, which approximately reflect the frequency of assessments per year (never = 0, 1-2 times a year = 1.5, 3-5 times a year = 4, monthly = 8, more than once a month = 12). The index of use of assessments (ASSESS) is calculated as the sum of these recoded items and then divided into three categories (less than 20 times a year, 20-39 times a year, more than 40 times a year).

Ability grouping

To determine the amount of within-school ability grouping, school principals were asked to report the extent to which their school organises instruction differently for students with different abilities regarding the following policies and practices: i) mathematics classes studying similar content, but at different levels of difficulty; and ii) different classes studying different content or sets of mathematics topics that have different levels of difficulty. The index of ability grouping between classes (ABGROUP) was derived from these items by assigning schools to three categories: (1) schools with no ability grouping between any classes; (2) schools with one of these forms of ability grouping between classes for some classes; and (3) schools with one of these forms of ability grouping for all classes.

Mathematics activities at school

School principals were asked to report what activities to promote engagement with mathematics occurred at their school; the list of activities included remedial or enrichment courses as well as other mathematics activities. The index of school offering extension courses (EXCOURSE) is computed as the sum of extension course types offered at school (none, either remedial or enrichment, both); the index of mathematics activity at school (MACTIV) is computed by simply counting the number of different activities occurring at school.

School management

School principals were asked to report whether teachers, department heads, the school principal, an appointed or elected board or education authorities at a higher level had the main responsibility for: i) selecting teachers for hire; ii) firing teachers; iii) establishing teachers' starting salaries; iv) determining teachers' salary increases; v) formulating school budgets; vi) deciding on budgets allocations within the school; vii) establishing student disciplinary policies; viii) establishing student assessment policies; ix) approving students for admittance to school; x) choosing which textbooks to use; xi) determining course content; and xii) deciding which courses are offered. The index of resource autonomy (AUTRES) is the number of decisions that relate to school resources which are a school responsibility (items i) to vi)), the index of curricular autonomy (AUTCURR) is the number of decisions that relate to curriculum which are a school responsibility (items viii), x), xi) and xii)). Two additional indices on (overall) school autonomy and teacher participation are described in the section on scaled indices below.

Poor student-teacher relations

An index of poor student-teacher relations at school (MSTREL) was derived from student responses to five items: i) most teachers are interested in students' well-being; ii) students who need extra help will receive it from their teacher; iii) most teachers treat students fairly; iv) students get along well with most teachers; and, v) most teachers really listen to what students have to say. The four-point scale with the response categories "strongly agree", "agree", "disagree" and "strongly disagree" was recoded into binary variables with strongly disagree coded 1 and other valid responses coded 0. These responses were summarised by taking the average item response per student and computing the mean for each school. See also the description of the student-level index on student-teacher relations in the section on scaled indices below.

METHODOLOGY

Scaling procedures

Most questionnaire items were scaled using IRT (Item Response Theory) scaling methodology. With the One-Parameter (Rasch) model (Rasch, 1960) for dichotomous items, the probability of selecting category 1 instead of 0 is modelled as

$$P_i(\theta) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \qquad (17.1)$$

where Pi(θ) is the probability of person n to score 1 on item i, θ_n is the estimated latent trait of person n and δi the estimated location of item i on this dimension. For each item, item responses are modelled as a function of the latent trait θ_n.

In the case of items with more than two (k) categories (as for example with Likert-type items) this model can be generalised to the partial credit model (Masters and Wright, 1997),[3] which takes the form of

$$P_{x_i}(\theta) = \frac{\exp(\sum_{j=0}^{x} \theta_n - \delta_i + \tau_{ij})}{1 + \exp(\sum_{j=1}^{k} \theta_n - \delta_i + \tau_{ij})} \qquad x_i = 0,1,\ldots,m_i \qquad (17.2)$$

where $P_{x_i}(\theta)$ denotes the probability of person n to score x on item i. θ_n denotes the person's latent trait, the item parameter δi gives the location of the item on the latent continuum and τ_{ij} denotes an additional step parameter.

Item fit was assessed using the weighted mean-square statistic (infit), which is a residual-based fit statistic. Weighted infit statistics were reviewed both for item and step parameters. The ACER ConQuest software (Wu *et al.,* 1997) was used for the estimation of item parameters and the analysis of item fit.

International item parameters were obtained from calibration samples consisting of randomly selected sub-samples:

- For the calibration of student item parameters, sub-samples of 500 students were randomly selected within each OECD country sample. As final student weights were not available at the time the calibration sample was drawn, the random selection was based on preliminary student weights obtained from the ratio between sampled and enrolled student within explicit sampling strata. The final calibration sample included data from 15 000 students.
- For the calibration of school item parameters, 100 schools were randomly selected within each OECD country sample. The random selection was based on school-level weights in order to ensure that a representative sample of schools was selected from each country. School data from Luxembourg were not included due to of the small number of schools. Data from France were not included because the school questionnaire had not been administered in France. The final calibration sample included data from 2 800 school principals.

Once international item parameters had been estimated from the calibration sample, weighted likelihood estimation was used to obtain individual student or school scores. Weighted Likelihood Estimates (WLEs) can be computed by solving the equation

$$\sum_{i\in\Omega}\left[\left(r_x+\frac{J_n}{2I_n}\right)-\sum_{j=1}^{k}\frac{\exp(\sum_{j=0}^{x}\theta_n-\delta_i+\tau_{ij})}{1+\exp(\sum_{j=1}^{k}\theta_n-\delta_i+\tau_{ij})}\right]=0 \quad (17.3)$$

for each case n, where r_x is the sum score obtained from a set of k items with j categories. This can be achieved by applying the Newton-Raphson method. The term $J_n/2I_n$ (with I_n being the information function for case n and J_n being its derivative with respect to θ) is used as a weight function to account for the bias inherent in maximum likelihood estimation (Warm, 1989). IRT scores were derived using an SPSS macro specifically designed for computing WLEs with pre-calibrated item parameters.

WLEs were transformed to an international metric with an OECD average of zero and an OECD standard deviation of one. The transformation was achieved by applying the formula

$$\theta'_n=\frac{\theta_n-\bar{\theta}_{OECD}}{\sigma_{\theta\,(OECD)}} \quad (17.4)$$

where θ'_n are the scores in the international metric, θ_n the original WLEs in logits, and $\bar{\theta}_{OECD}$ is the OECD mean of logit scores with equally weighted country sub-samples. $\sigma_{\theta\,(OECD)}$ is the corresponding OECD standard deviation of the original WL estimates. School scores were standardised using student-level weights. Means and standard deviations used for the transformation into the international metric are shown in Table 17.1.

Describing questionnaire scale indices

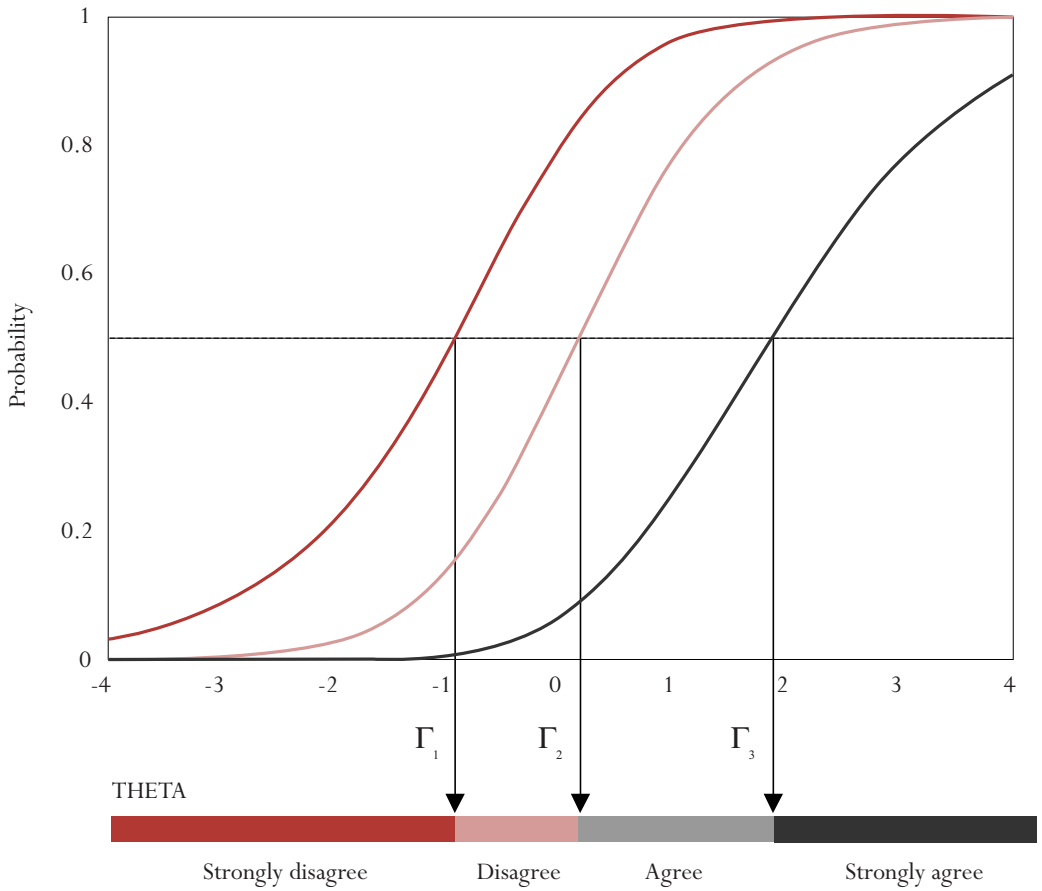
In PISA 2003 categorical items from the context questionnaires were scaled using IRT modelling. WLEs (logits) for the latent dimensions were transformed to scales with an OECD average of zero and a standard deviation of one (with equally weighted samples). It is possible to interpret these scores by comparing individual scores or group average scores to the OECD mean, but the individual scores do not reveal anything about the actual item responses and it is impossible to determine from scale score values to what extent respondents endorsed the items used for the measurement of the latent variable. However, the scaling model used to derive individual scores allows the provision of descriptions of these scales by mapping scale scores to (expected) item responses.[4]

Item characteristics can be described, using the parameters of the partial credit model, by summing, for each category, its

Table 17.1 ■ **OECD means and standard deviations of WLEs for indices**

Student-level indices	Mean	Standard deviation
ANXMAT	-0.34	1.81
ATSCHL	1.00	1.34
ATTCOMP	2.02	1.88
BELONG	1.81	1.78
COMPHOME	0.66	2.03
COMPLRN	0.27	2.05
COOPLRN	0.51	1.69
CSTRAT	0.98	1.56
CULTPOSS	-0.05	1.49
DISCLIM	0.65	1.69
ELAB	0.00	1.48
HEDRES	1.98	1.09
HIGHCONF	0.93	1.47
HOMEPOS	1.25	1.46
INSTMOT	1.16	2.59
INTCONF	2.61	1.64
INTMAT	-0.94	2.89
INTUSE	0.09	1.05
MATHEFF	1.00	1.39
MEMOR	0.13	1.22
PRGUSE	-0.54	1.04
ROUTCONF	2.87	1.30
SCMAT	-0.30	2.40
STUREL	0.61	1.81
TEACHSUP	0.68	1.65
School-level indices		
SCHAUTON	1.45	2.12
SCMATBUI	0.62	1.84
SCMATEDU	0.79	1.66
STMORALE	0.12	2.25
STUDBEHA	1.07	1.69
TCHCONS	2.07	2.55
TCHPARTI	-2.77	1.95
TCMORALE	1.29	2.26
TCSHORT	-1.56	1.72
TEACBEHA	1.44	1.56

Figure 17.1 ■ **Summed category probabilities for a fictitious item**



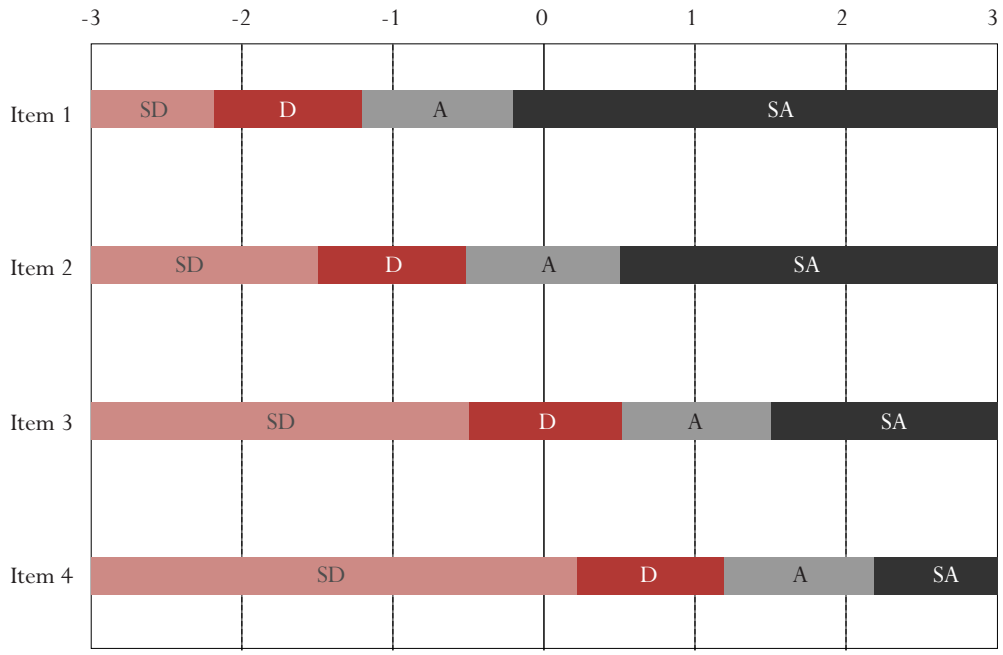
probability of being chosen with the probabilities of all higher categories. This is equivalent to computing the odds of scoring higher than a particular category.

The results of plotting these cumulative probabilities against scale scores for a fictitious item are displayed in Figure 17.1. The three vertical lines denote those points on the latent continuum where it becomes more likely to score >0, >1 or >2. These locations Γk, called Thurstonian thresholds, can be obtained through an iterative procedure that calculates summed probabilities for each category at each (decimal) point on the latent variable.

Summed probabilities are not identical with expected item scores and have to be understood in terms of the odds to score at least a particular category. Other ways of describing the item characteristics based on the partial credit model are Item Characteristic Curves (by plotting the individual category probabilities) and Expected Item Score Curves (for a more detailed description see Masters and Wright, 1997).

Thurstonian thresholds can be used to indicate for each item category those points on a scale, at which respondents have a 0.5 probability of scoring this category or higher. For example, in the case of Likert-type items with categories strongly disagree (SD), disagree (D), agree (A) and strongly agree (SA), it is possible to determine at what point of a scale a respondent has 50 per cent chance of at least agreeing with the item.

Figure 17.2 ■ **Fictitious example of an item map**



The fictitious example in Figure 17.2 illustrates the interpretation of an item map for a fictitious student questionnaire scale with four different Likert-type items:

- Students with a score of -2 (that is, 2 standard deviations below the OECD average) have more than a 0.5 probability to disagree, agree or strongly agree (or not to disagree strongly) with item 1, but they have more than a 50 per cent chance to strongly disagree with the other three items.
- Students with a score of -1 (one standard deviation below the OECD average), have already more than 0.5 probability to agree with the first item, but they would still be expected to disagree with item 2 or even to strongly disagree with item 3 and 4.
- Likewise, students with a score of 1 (one standard deviation above the OECD average) would have more than a 0.5 probability to strongly agree with the first two items, but still have less than a 0.5 probability to agree with item 4.

Item maps can help to illustrate the relationship between scores and item responses. For example, even scores of one standard deviation below the OECD average on an attitudinal scale could still indicate affirmative responses. This would not be revealed by the international metric, which has to be interpreted relative to the OECD average, but is illustrated by the corresponding item map.

Construct validation

One of the important challenges of international educational research is the search for comparable measures of student background, attitudes and perceptions. There are different methodological approaches for validating questionnaire constructs, each with their limitations and problems. Cross-country validation of these constructs is of particular importance, as measures derived from questionnaires are often used to

explain differences in student performance within and across countries and are, thus, potential sources of policy-relevant information about ways of improving educational systems.

Cross-country validation of the constructs requires a thorough and closely monitored process of translation into different languages. It also makes assumptions about having measured similar characteristics, attitudes and perceptions in different national and cultural contexts. Psychometric techniques can be used to analyse the extent to which constructs have consistent dimensionality and consistent construct validity across participating countries. This means that, once the measurement stability for each scale is confirmed, the multidimensional relationship between these constructs should be reviewed as well (see Wilson, 1994; Schulz 2003). It should be noted, however, that between-country differences in the strength of relationships between constructs do not necessarily indicate a lack of consistency, as they may be due to differences between national contexts (for example, different educational systems or learning practices).

Confirmatory factor analysis

Structural Equation Modelling (SEM) was used to confirm theoretically expected dimensions and, if necessary, to re-specify the dimensional structure. Using Confirmatory Factor Analysis (CFA) requires a theoretical model of item dimensionality, which can be tested using the collected data (see Kaplan, 2000).

Fit indices measure the extent to which a model based on the a-priori structure as postulated by the analyst fits the data. In the PISA 2003 analysis, model fit was assessed using the Root-Mean Square Error of Approximation (RMSEA), the Root Mean Square Residual (RMR), the Comparative Fit Index (CFI) and the Non-normed Fit Index (NNFI) (see Bollen and Long, 1993). RMSEA values over 0.10 are usually interpreted as a sign of unacceptable model fit, whereas values below 0.05 indicate a close model fit. RMR values should be less than 0.05. Both CFI and NNFI are bound between 0 and 1. Values between 0.90 and 0.95 indicate an acceptable model fit, values greater than 0.95 indicate a close model fit.

Generally, maximum likelihood estimation and covariance matrices were used for the analyses of the (categorical) Likert-type items, that is, the items were treated as if they were continuous. Confirmatory factor analyses of student data were based on the international calibration sample, in order to have comparable (sub-)sample sizes across OECD countries. For the comparative analysis of item dimensionality, the use of random OECD sub-samples was deemed appropriate.

The SAS CALIS procedure (Hatcher, 1994) and the LISREL programme (Jöreskog and Sörbom, 1993) with the STREAMS interface programme (Gustafson 2000) were used to estimate the models based on Likert-type items. In order to assess cross-country validity of item dimensionality, and constructs, models were estimated both for the international pooled sample and for country sub-samples separately.

In addition, multiple group models were estimated which allow researchers to test the invariance of parameters across sub-samples (in this case, countries). Series of different models from least restrictive (with different parameters for each group) to most restrictive models (with all parameters being the same across groups) can be tested (see Marsh, 1994). Within the context of the PISA 2003 cross-country validation, only the minimal condition for factorial invariance, the equivalence of factor loadings across countries, was tested. The relationships between PISA 2003 constructs are likely to be influenced by the context and structure of the educational systems. Therefore, testing invariance of relations between constructs is of interest but not a necessary condition for cross-country validity.

In the case of dichotomous items, Weighted Least Squares (WLS) estimation with polychoric correlations was used (Jöreskog and Sörbom, 1993). As the unadjusted WLS estimator requires very large sample sizes, a mean- and variance-adjusted WLS estimator (WLSMV) was used, which is available in the Mplus software program, (Muthén *et al.*, 1997, and Muthén and Muthén, 2003). The confirmatory factor analyses for dichotomous student-level items were only estimated for the pooled international calibration sample.

Confirmatory factor analyses of school-level data were based on the international school calibration sample, with 100 schools per OECD country.

QUESTIONNAIRE SCALE INDICES

Student indices

Household possessions

Using data about household possessions as an indicator of family wealth has received much attention in recent international studies in the field of education (Buchmann, 2000). Data about household assets are believed to capture wealth better than income because they reflect a more stable source of wealth. In PISA 2003, students reported the availability of 13 different household items at home. Four different indices were derived from these items: computer facilities at home (COMPHOME); cultural possessions (CULTPOSS); home educational resources (HEDRES); and home possessions (HOMEPOS). The last index is a summary index of all household items and also included a dummy variable indicating more than 100 books (derived from a question on the number of books at home). It was also one of three components in the construction of the index on economic, social and cultural status (ESCS, see the section on ESCS index construction below). Table 17.2 shows the wording of items and the IRT model parameters for the four indices.

Table 17.2 ■ **Item parameters for home background indices**

		Item parameters for scale…			
	In your home, do you have:	**COMPHOME**	**CULTPOSS**	**HEDRES**	**HOMEPOS**
ST17Q01	a) A desk for study			-0.26	-1.66
ST17Q02	b) A room of your own				-0.88
ST17Q03	c) A quiet place to study			0.78	-0.70
ST17Q04	d) A computer you can use for school work	-1.58			-0.42
ST17Q05	e) Educational software	1.65			1.62
ST17Q06	f) A link to the Internet	-0.08			0.51
ST17Q07	g) Your own calculator			-0.46	-1.87
ST17Q08	h) Classic literature (*e.g.* <author>)		-0.08		1.29
ST17Q09	i) Books of poetry		0.07		1.40
ST17Q10	j) Works of art (*e.g.* paintings)		0.01		1.35
ST17Q11	k) Books to help with your school work			1.01	-0.49
ST17Q12	l) A dictionary			-1.07	-2.43
ST17Q13	l) A dishwasher				0.75
	In your home, do you have:				
ST19Q01	More than 100 books (recoded)				1.54

Note: Item categories were "yes" and "no" and all items were inverted for scaling.

A confirmatory factor analysis using polychoric correlations with a WLSMV estimator showed a reasonable model fit for the international calibration sample (RMSEA = 0.066, CFI = 0.95, NNFI = 0.95). The estimated latent correlations between these constructs were 0.37 between COMPHOME and CULTPOSS, 0.55 between COMPHOME and HEDRES, and 0.75 between CULTPOSS and HEDRES.[5]

Some of the item reliabilities tended to be very low for the index of home educational resources (HEDRES). This is also reflected in the lower scale reliabilities for this index (see Table 17.3 with the scale reliabilities for all four indices). Similar results for this index were found in PISA 2000 (see OECD, 2002).

Table 17.3 ■ **Reliabilities for home background indices**

		COMPHOME	CULTPOS	HEDRES	HOMEPOS
OECD countries	Australia	0.58	0.71	0.57	0.75
	Austria	0.43	0.67	0.42	0.71
	Belgium	0.65	0.69	0.57	0.75
	Canada	0.59	0.69	0.49	0.72
	Czech Republic	0.73	0.64	0.47	0.72
	Denmark	0.45	0.71	0.56	0.76
	Finland	0.60	0.71	0.54	0.73
	France	0.67	0.67	0.46	0.73
	Germany	0.54	0.68	0.55	0.73
	Greece	0.70	0.59	0.43	0.73
	Hungary	0.64	0.66	0.49	0.72
	Iceland	0.44	0.68	0.48	0.71
	Ireland	0.70	0.67	0.54	0.74
	Italy	0.68	0.67	0.59	0.75
	Japan	0.61	0.65	0.50	0.70
	Korea	0.42	0.66	0.48	0.75
	Luxembourg	0.54	0.70	0.50	0.73
	Mexico	0.81	0.65	0.64	0.79
	Netherlands	0.43	0.63	0.50	0.72
	New Zealand	0.72	0.66	0.58	0.76
	Norway	0.47	0.77	0.58	0.74
	Poland	0.82	0.58	0.58	0.76
	Portugal	0.70	0.68	0.51	0.78
	Slovak Republic	0.66	0.64	0.58	0.71
	Spain	0.63	0.65	0.47	0.72
	Sweden	0.50	0.72	0.65	0.78
	Switzerland	0.52	0.67	0.56	0.73
	Turkey	0.78	0.63	0.63	0.80
	United Kingdom	0.66	0.74	0.63	0.77
	United States	0.73	0.73	0.63	0.81
	OECD median	0.63	0.67	0.54	0.74
Partner countries	Brazil	0.81	0.57	0.56	0.75
	Hong Kong-China	0.49	0.59	0.55	0.70
	Indonesia	0.54	0.57	0.55	0.65
	Latvia	0.77	0.65	0.54	0.70
	Liechtenstein	0.40	0.71	0.50	0.70
	Macao-China	0.51	0.57	0.52	0.69
	Russian Federation	0.84	0.62	0.58	0.74
	Serbia	0.81	0.74	0.62	0.77
	Thailand	0.84	0.61	0.58	0.75
	Tunisia	0.66	0.56	0.64	0.77
	Uruguay	0.86	0.61	0.52	0.75

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

School climate indices

Three student-level indices related to the school climate were derived. Two of these constructs, student-teacher relations at school (STUREL) and sense of belonging at school (BELONG), were included in PISA 2000 (OECD, 2003). In addition, four items on attitudes towards school (ATSCHL) were administered to provide data on general attitudes of students towards schooling.

Table 17.4 shows item wording and IRT parameters for items measuring students' attitudes towards school. Two of these items are phrased positively and were inverted for scaling, so that positive scores indicate positive attitudes towards school.

Table 17.4 ■ **Item parameters for attitudes towards school (ATSCHL)**

	Thinking about what you have learned in school: To what extent do you agree with the following statements?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST24Q01	a) School has done little to prepare me for adult life when I leave school.	0.48	-1.72	-0.37	2.08
ST24Q02	b) School has been a waste of time.	-0.59	-0.47	-1.19	1.66
ST24Q03	c) School helped give me confidence to make decisions. [a]	0.50	-1.55	-0.73	2.28
ST24Q04	d) School has taught me things which could be useful in a job. [a]	-0.39	-0.77	-0.85	1.62

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree".
a. Item inverted for scaling.

Five items from PISA 2000 were also included in the PISA 2003 student questionnaire to measure students' perceptions of student-teacher relations. All items were inverted for scaling so that positive scores indicate positive perceptions of student-teacher relations at school. Table 17.5 illustrates item wording and IRT parameters for this index.

Table 17.5 ■ **Item parameters for student-teacher relations at school (STUREL)**

	Thinking about the teachers at your school: To what extent do you agree with the following statements?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST26Q01	a) Students get along well with most teachers.	0.07	-2.73	-0.77	3.50
ST26Q02	b) Most teachers are interested in students' well-being.	0.15	-2.72	-0.63	3.35
ST26Q03	c) Most of my teachers really listen to what I have to say.	0.21	-2.85	-0.48	3.33
ST26Q04	d) If I need extra help, I will receive it from my teachers.	-0.26	-2.42	-0.81	3.23
ST26Q05	e) Most of my teachers treat me fairly.	-0.17	-2.25	-1.04	3.29

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

Six items from PISA 2000 were retained for the PISA 2003 assessment to measure students' perceptions of school. The three positively phrased items were inverted for scaling so that positive scores indicate positive feelings about the students' school. Table 17.6 shows item wording and IRT parameters for this index.

Table 17.6 ■ **Item parameters for sense of belonging to school (BELONG)**

	My school is a place where:	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST27Q01	a) I feel like an outsider (or left out of things).	-0.50	-1.20	-0.92	2.12
ST27Q02	b) I make friends easily.[a]	0.03	-2.11	-1.04	3.14
ST27Q03	c) I feel like I belong.[a]	0.69	-1.90	-0.93	2.83
ST27Q04	d) I feel awkward and out of place.	-0.26	-1.47	-0.84	2.31
ST27Q05	e) Other students seem to like me.[a]	0.48	-2.31	-1.39	3.69
ST27Q06	f) I feel lonely.	-0.44	-0.94	-0.83	1.78

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree".
a. Item inverted for scaling.

Results from a confirmatory factor analysis of all items measuring school climate factors support the construct validity of these indices. Model fit was good for the international sample as well across country sub-samples (see tables 17.1 and 17.2).[6] The estimated latent correlations between the constructs show that correlation is highest for ATSCHL and STUREL.

A comparison between the unrestricted multiple-group model (RMSEA = 0.068, RMR = 0.020, CFI = 0.90, NNFI = 0.88) and the model with constrained factor loadings (RMSEA = 0.069, RMR = 0.022, CFI = 0.89, NNFI = 0.88) shows a high degree of invariance for these parameters and provides support for the cross-country validity of this model.

Table 17.8 shows that the internal consistency was only moderate to poor for ATSCHL but good for the other two indices.

Motivations in mathematics

Subject-related interest is an intrinsic motivational preference, which affects continuity and intensity of engagement with learning, independently of the general motivation to learn (for an overview on interest research, see Baumert and Koeller, 1998). Closely related to the interest dimension are enjoyment of mathematics and value of mathematics (Aiken 1974). Intrinsic motivation is viewed as having positive effects on: time on task, more comprehensive learning strategies, performance and activity choices in the absence of extrinsic rewards (Lepper, 1988). There is evidence suggesting that intrinsic motivation to learn is at least partially influenced by teacher supportiveness and classroom environment (Middleton and Spanias, 1999).

Four items were used to measure interest in and enjoyment of mathematics in PISA 2003, all of them had been tested in the field trial, with only item a) subsequently modified (reading instead of reading books). All items were inverted for scaling, and positive scores indicate higher levels of interest in and enjoyment of mathematics. Item wording and model parameters are displayed in Table 17.9.

Table 17.7 ■ **Model fit and estimated latent correlations for school climate items**

Country[1]	Model fit				Latent correlations between:		
	RMSEA	RMR	CFI	NNFI	ATSCHL/ STUREL	ATSCHL/ BELONG	STUREL/ BELONG
Australia	0.085	0.023	0.88	0.88	0.63	0.30	0.34
Austria	0.067	0.030	0.89	0.89	0.57	0.19	0.16
Belgium	0.068	0.032	0.89	0.89	0.67	0.43	0.37
Canada	0.054	0.019	0.95	0.95	0.48	0.39	0.16
Czech Republic	0.053	0.022	0.92	0.92	0.59	0.42	0.28
Denmark	0.065	0.039	0.91	0.91	0.73	0.41	0.26
Finland	0.073	0.017	0.92	0.92	0.55	0.32	0.29
France	0.047	0.028	0.93	0.93	0.58	0.25	0.19
Germany	0.080	0.037	0.83	0.83	0.58	0.29	0.26
Greece	0.061	0.027	0.91	0.91	0.58	0.38	0.24
Hungary	0.054	0.021	0.92	0.92	0.50	0.26	0.18
Iceland	0.087	0.030	0.89	0.89	0.65	0.32	0.20
Ireland	0.066	0.024	0.91	0.91	0.64	0.37	0.15
Italy	0.060	0.023	0.91	0.91	0.60	0.43	0.22
Japan	0.077	0.043	0.86	0.87	0.44	0.45	0.42
Korea	0.057	0.022	0.93	0.93	0.48	0.34	0.21
Luxembourg	0.053	0.028	0.92	0.92	0.68	0.21	0.21
Mexico	0.073	0.028	0.85	0.85	0.36	0.43	0.42
Netherlands	0.057	0.018	0.90	0.90	0.65	0.43	0.13
New Zealand	0.055	0.020	0.94	0.94	0.58	0.35	0.25
Norway	0.070	0.028	0.91	0.91	0.81	0.35	0.25
Poland	0.060	0.030	0.89	0.89	0.53	0.18	0.19
Portugal	0.067	0.021	0.87	0.87	0.38	0.40	0.14
Slovak Republic	0.072	0.021	0.86	0.86	0.53	0.45	0.28
Spain	0.065	0.025	0.89	0.89	0.53	0.33	0.17
Sweden	0.066	0.026	0.91	0.91	0.59	0.39	0.22
Switzerland	0.059	0.028	0.93	0.93	0.63	0.53	0.40
Turkey	0.059	0.043	0.85	0.86	0.55	0.25	0.25
United Kingdom	0.069	0.020	0.93	0.93	0.65	0.33	0.27
United States[2]	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OECD	**0.053**	**0.017**	**0.93**	**0.92**	**0.61**	**0.37**	**0.26**

1. Model estimates based on international student calibration sample (500 students per OECD country).
2. United States did not administer the items measuring BELONG so that their data could not be included in these analyses.

Table 17.8 ■ **Reliabilities for school climate indices**

		ATTSCH	STUREL	BELONG
OECD countries	Australia	0.70	0.83	0.85
	Austria	0.61	0.77	0.81
	Belgium	0.58	0.75	0.76
	Canada	0.70	0.80	0.85
	Czech Republic	0.57	0.77	0.73
	Denmark	0.62	0.81	0.78
	Finland	0.68	0.80	0.86
	France	0.63	0.70	0.76
	Germany	0.54	0.76	0.81
	Greece	0.59	0.74	0.76
	Hungary	0.55	0.77	0.79
	Iceland	0.72	0.83	0.86
	Ireland	0.69	0.80	0.82
	Italy	0.63	0.76	0.79
	Japan	0.65	0.81	0.76
	Luxembourg	0.58	0.80	0.78
	Mexico	0.46	0.75	0.72
	Netherlands	0.50	0.73	0.75
	New Zealand	0.69	0.81	0.83
	Norway	0.68	0.82	0.83
	Poland	0.61	0.74	0.75
	Portugal	0.60	0.75	0.78
	Republic of Korea	0.72	0.77	0.75
	Slovak Republic	0.61	0.74	0.77
	Spain	0.60	0.77	0.79
	Sweden	0.64	0.81	0.82
	Switzerland	0.59	0.82	0.77
	Turkey	0.50	0.69	0.55
	United Kingdom	0.71	0.82	0.84
	United States [1]	0.68	0.82	N/A
	OECD median	0.58	0.76	0.74
Partner countries	Brazil	0.54	0.73	0.80
	Hong Kong-China	0.65	0.80	0.74
	Indonesia	0.45	0.62	0.55
	Latvia	0.60	0.73	0.74
	Liechtenstein	0.58	0.83	0.82
	Macao-China	0.66	0.78	0.75
	Russian Federation	0.62	0.68	0.73
	Serbia	0.62	0.80	0.73
	Thailand	0.44	0.76	0.70
	Tunisia	0.50	0.76	0.69
	Uruguay	0.53	0.76	0.79

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.
1. United States did not administer the items measuring BELONG.

Table 17.9 ■ **Item parameters for interest in and enjoyment of mathematics (INTMAT)**

	Thinking about your views on mathematics: To what extent do you agree with the following statements?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST30Q01	a) I enjoy reading about mathematics.	0.61	-3.60	0.08	3.52
ST30Q03	c) I look forward to my mathematics lessons.	0.48	-3.68	0.11	3.57
ST30Q04	d) I do mathematics because I enjoy it.	-0.10	-3.40	0.11	3.29
ST30Q06	f) I am interested in the things I learn in mathematics.	-0.99	-3.77	-0.29	4.06

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

Instrumental motivation has been found to be an important predictor for course selection, career choice and performance (Eccles, 1994; Eccles and Wigfield, 1995; Wigfield *et al.,* 1998).

Four items measuring instrumental motivation were used in the main study of PISA 2003. All items were inverted for scaling and positive scores on this index indicate higher levels of instrumental motivation to learn mathematics. Table 17.10 shows the item wording and the model parameters used for IRT scaling.

Table 17.10 ■ **Item parameters for instrumental motivation to learn mathematics (INSTMOT)**

	Thinking about your views on mathematics: To what extent do you agree with the following statements?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST30Q02	b) Making an effort in mathematics is worth it because it will help me in the work that I want to do later on.	-0.25	-2.86	-0.67	3.52
ST30Q05	e) Learning mathematics is worthwhile for me because it will improve my career <prospects, chances>.	-0.31	-2.67	-0.91	3.58
ST30Q07	g) Mathematics is an important subject for me because I need it for what I want to study later on.	0.35	-2.61	-0.48	3.09
ST30Q08	h) I will learn many things in mathematics that will help me get a job.	0.20	-2.97	-0.63	3.60

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

The fit for a two-factor model was satisfactory for the pooled international sample and was acceptable in all but two country sub-samples (Table 17.11). The results show a relatively high correlation between the two constructs and the strength of that correlation does not vary much across country sub-samples.

When comparing the multiple-group models, the fit indices with constrained factor loadings (RMSEA = 0.077, RMR = 0.036, CFI = 0.97, NNFI = 0.97) is only slightly different from those with unconstrained factor loadings (RMSEA = 0.078, RMR = 0.026, CFI = 0.98, NNFI = 0.96). This provides further support for the parameter invariance of the model across countries.

Table 17.12 shows that both indices have a remarkable degree of internal consistency across participating countries.

Table 17.11 ■ **Model fit and estimated latent correlations for motivation items**

	Model fit				Latent correlations between:
	RMSEA	RMR	CFI	NNFI	INTMAT/ INSTMOT
Australia	0.069	0.023	0.98	0.98	0.61
Austria	0.055	0.027	0.98	0.98	0.60
Belgium	0.086	0.031	0.97	0.95	0.73
Canada	0.096	0.031	0.97	0.95	0.67
Czech Republic	0.067	0.021	0.98	0.97	0.59
Denmark	0.023	0.015	1.00	1.00	0.67
Finland	0.072	0.025	0.98	0.97	0.64
France	0.081	0.031	0.97	0.95	0.70
Germany	0.085	0.047	0.97	0.95	0.56
Greece	0.095	0.045	0.97	0.95	0.66
Hungary	0.029	0.012	1.00	0.99	0.67
Iceland	0.046	0.016	0.99	0.99	0.68
Ireland	0.074	0.026	0.98	0.97	0.60
Italy	0.091	0.026	0.97	0.95	0.68
Japan	0.099	0.034	0.97	0.96	0.68
Korea	0.054	0.016	0.99	0.99	0.72
Luxembourg	0.056	0.025	0.99	0.98	0.65
Mexico	0.094	0.025	0.94	0.92	0.67
Netherlands	0.068	0.024	0.98	0.97	0.59
New Zealand	0.073	0.026	0.98	0.97	0.50
Norway	0.112	0.046	0.96	0.94	0.68
Poland	0.063	0.018	0.98	0.98	0.70
Portugal	0.072	0.025	0.98	0.96	0.65
Slovak Republic	0.074	0.023	0.97	0.96	0.69
Spain	0.121	0.045	0.95	0.92	0.72
Sweden	0.077	0.028	0.98	0.96	0.59
Switzerland	0.098	0.036	0.96	0.95	0.71
Turkey	0.084	0.036	0.97	0.96	0.68
United Kingdom	0.060	0.023	0.99	0.98	0.62
United States	0.070	0.026	0.98	0.98	0.66
OECD	**0.053**	**0.017**	**0.93**	**0.92**	**0.61**

Note: Model estimates based on international student calibration sample (500 students per OECD country).

Self-related cognitions in mathematics

Bandura (1986) stated that self-efficacy plays an important role in determining behaviour and that feelings of confidence about a specific problem are crucial to an individual's capacity to solve that problem. Research has generally confirmed a relationship between mathematics self-efficacy and student performance, although different sizes of correlation were reported, often depending on the types of self-efficacy measures that were used (Multon *et al.,* 1991). It has been found that task-specific mathematics self-efficacy was a better predictor of career choice than test performance (Hackett and Betz, 1989).

Eight items measuring the students' confidence with mathematical tasks had been tested in the field trial and were (after minor modifications) retained for the main study in 2003. All items were Table 17.13 shows the item wording and the model parameters used for IRT scaling.

Mathematics anxiety is concerned with feelings of helplessness and emotional stress when dealing with mathematics. Mathematics anxiety is usually found to be negatively associated with achievement but this relationship can change depending on the students' social and academic background (Ma, 1999). It could also be shown that mathematics anxiety has rather indirect effects on achievement, once self-related cognitions such as self-efficacy and self-concept are taken into account (Meece *et al.,* 1990).

Six items measuring mathematics anxiety were piloted in the field trial, five of them were retained for the main study. All items were inverted for scaling and positive scores indicate higher levels of mathematics anxiety. Item wording and model parameters are shown in Table 17.14.

Table 17.12 ■ **Reliabilities for indices of motivation in mathematics**

		INTMAT	INSTMOT
OECD countries	Australia	0.90	0.89
	Austria	0.89	0.83
	Belgium	0.88	0.88
	Canada	0.91	0.90
	Czech Republic	0.85	0.85
	Denmark	0.91	0.84
	Finland	0.90	0.88
	France	0.87	0.87
	Germany	0.90	0.82
	Greece	0.90	0.88
	Hungary	0.88	0.85
	Iceland	0.92	0.89
	Ireland	0.90	0.87
	Italy	0.88	0.87
	Japan	0.90	0.91
	Korea	0.91	0.88
	Luxembourg	0.89	0.90
	Mexico	0.82	0.77
	Netherlands	0.88	0.87
	New Zealand	0.90	0.88
	Norway	0.91	0.88
	Poland	0.89	0.85
	Portugal	0.83	0.89
	Slovak Republic	0.85	0.86
	Spain	0.88	0.89
	Sweden	0.91	0.86
	Switzerland	0.88	0.85
	Turkey	0.90	0.84
	United Kingdom	0.90	0.86
	United States	0.91	0.89
	OECD median	0.90	0.87
Partner countries	Brazil	0.84	0.84
	Hong Kong-China	0.91	0.88
	Indonesia	0.83	0.77
	Latvia	0.82	0.84
	Liechtenstein	0.88	0.87
	Macao-China	0.88	0.85
	Russian Federation	0.86	0.87
	Serbia	0.87	0.87
	Thailand	0.85	0.81
	Tunisia	0.87	0.86
	Uruguay	0.89	0.87

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

Table 17.13 ■ **Item parameters for mathematics self-efficacy (MATHEFF)**

	How confident do you feel about having to do the following calculations?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST31Q01	a) Using a <train timetable>, how long it would take to get from Zedville to Zedtown	-0.27	-1.72	-0.15	1.88
ST31Q02	b) Calculating how much cheaper a TV would be after a 30 percent discount	-0.36	-1.61	-0.01	1.62
ST31Q03	c) Calculating how many square metres of tiles you need to cover a floor	0.07	-1.74	0.17	1.57
ST31Q04	d) Understanding graphs presented in newspapers	-0.17	-1.58	-0.24	1.82
ST31Q05	e) Solving an equation like $3x + 5 = 17$	-0.56	-0.93	-0.02	0.95
ST31Q06	f) Finding the actual distance between two places on a map with a 1:10,000 scale	0.50	-1.82	0.24	1.58
ST31Q07	g) Solving an equation like $2(x+3) = (x + 3)(x - 3)$	0.12	-1.23	0.03	1.20
ST31Q08	h) Calculating the petrol consumption rate of a car	0.68	-1.93	0.10	1.83

Note: Categories were "very confident", "confident", "not very confident" and "not at all confident". All items were inverted for scaling.

Table 17.14 ■ **Item parameters for mathematics anxiety (ANXMAT)**

	How much do you disagree or agree with the following statements about how you feel when studying mathematics?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST32Q01	a) I often worry that it will be difficult for me in mathematics classes.	-0.68	-2.39	-0.02	2.41
ST32Q03	c) I get very tense when I have to do mathematics homework.	0.44	-2.45	0.54	1.91
ST32Q05	e) I get very nervous doing mathematics problems.	0.45	-2.67	0.55	2.12
ST32Q08	h) I feel helpless when doing a mathematics problem.	0.48	-2.53	0.54	1.99
ST32Q10	j) I worry that I will get poor <marks> in mathematics.	-0.68	-1.77	-0.11	1.87

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

Positive self-concept can be seen as a desirable outcome variable of education (Branden, 1994), and its enhancement is one of the goals of policy makers. It can be assumed that students evaluate their own performance through social comparison processes, and it has been observed that the school average of achievement tends to have a negative effect on self-concept and that students with same proficiency levels often have different levels of self-concept depending on the overall performance of a school (Marsh, 1990).

Eight items on mathematics self-concept had been piloted in the international field trial and five were retained for the main study. One item was negatively phrased and therefore not inverted for scaling. All other items were inverted for scaling so that positive scores indicate a positive self-concept in mathematics. Table 17.15 shows the item wording and the IRT model parameters used for scaling.

Table 17.15 ■ Item parameters for mathematics self-concept (SCMAT)

	How much do you disagree or agree with the following statements about how you feel when studying mathematics?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST32Q02	b) I am just not good at mathematics.	-0.57	-2.42	-0.43	2.85
ST32Q04	d) I get good <marks> in mathematics.[1]	-0.52	-2.98	-0.33	3.30
ST32Q06	f) I learn mathematics quickly.[1]	-0.27	-3.12	-0.16	3.28
ST32Q07	g) I have always believed that mathematics is one of my best subjects.[1]	0.51	-2.41	0.28	2.13
ST32Q09	i) In my mathematics class, I understand even the most difficult work.[1]	0.84	-3.18	0.10	3.08

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree".
1. Item inverted for scaling.

Table 17.16 ■ Model fit and estimated latent correlations for self-related cognitions

	Model fit				Latent correlations between:		
	RMSEA	RMR	CFI	NNFI	MATHEFF/ ANXMAT	MATHEFF/ SCMAT	ANXMAT/ SCMAT
Australia	0.072	0.031	0.92	0.91	-0.55	0.72	-0.78
Austria	0.092	0.054	0.86	0.84	-0.42	0.48	-0.87
Belgium	0.075	0.041	0.91	0.90	-0.36	0.43	-0.82
Canada	0.082	0.042	0.92	0.91	-0.55	0.64	-0.91
Czech Republic	0.086	0.042	0.88	0.86	-0.57	0.61	-0.81
Denmark	0.071	0.031	0.93	0.92	-0.75	0.83	-0.91
Finland	0.065	0.031	0.94	0.93	-0.64	0.77	-0.82
France	0.088	0.060	0.87	0.85	-0.43	0.60	-0.75
Germany	0.085	0.043	0.90	0.89	-0.49	0.54	-0.89
Greece	0.088	0.056	0.85	0.83	-0.58	0.73	-0.87
Hungary	0.071	0.033	0.90	0.88	-0.43	0.40	-0.86
Iceland	0.105	0.070	0.88	0.86	-0.61	0.76	-0.70
Ireland	0.079	0.043	0.90	0.88	-0.59	0.67	-0.84
Italy	0.091	0.052	0.84	0.82	-0.41	0.56	-0.68
Japan	0.088	0.049	0.90	0.88	-0.45	0.47	-0.86
Korea	0.094	0.040	0.86	0.84	-0.49	0.69	-0.83
Luxembourg	0.071	0.045	0.91	0.90	-0.37	0.45	-0.78
Mexico	0.084	0.042	0.83	0.80	-0.43	0.52	-0.77
Netherlands	0.077	0.036	0.91	0.89	-0.54	0.61	-0.85
New Zealand	0.072	0.031	0.92	0.91	-0.64	0.72	-0.83
Norway	0.071	0.045	0.93	0.92	-0.62	0.77	-0.81
Poland	0.088	0.040	0.86	0.84	-0.57	0.66	-0.82
Portugal	0.090	0.043	0.87	0.85	-0.40	0.65	-0.72
Slovak Republic	0.070	0.030	0.92	0.91	-0.56	0.60	-0.87
Spain	0.090	0.053	0.86	0.84	-0.25	0.53	-0.63
Sweden	0.081	0.039	0.91	0.89	-0.59	0.77	-0.78
Switzerland	0.085	0.045	0.89	0.87	-0.56	0.61	-0.85
Turkey	0.089	0.056	0.88	0.86	-0.47	0.62	-0.80
United Kingdom	0.077	0.037	0.92	0.91	-0.60	0.73	-0.80
United States	0.096	0.048	0.88	0.86	-0.54	0.55	-0.86
OECD	**0.077**	**0.036**	**0.91**	**0.89**	**-0.52**	**0.62**	**-0.80**

Note: Model estimates based on international student calibration sample (500 students per OECD country).

Table 17.16 shows the results of a confirmatory factor analysis for a three-factor model. The model fit is satisfactory for the pooled international sample and for most country sub-samples. All three constructs are highly correlated. MATHEFF is negatively correlated with ANXMAT and positively correlated with SCMATH. ANXMAT and SCMATH have very high negative correlations around -0.80. The strengths of the estimated latent correlation are very similar across country sub-samples.

The comparison of multiple-group models shows that the fit for the model with constrained factor loadings (RMSEA = 0.088, RMR = 0.061, CFI = 0.88, NNFI = 0.87) is only slightly less satisfactory than the fit for the unrestricted model (RMSEA = 0.087, RMR = 0.048, CFI = 0.89, NNFI = 0.88). These results provide some evidence of parameter invariance across OECD countries. Table 17.17 also demonstrates the very good reliabilities for all three indices across participating countries.

Learning strategies in mathematics

Students may develop different types of learning strategies that shape their learning behaviour. Some main cognitive strategies are memorisation (*e.g.* learning key terms, repeated learning of material) and elaboration (*e.g.* making connections to related areas, thinking about alternative solutions). Control strategies are meta-cognitive strategies that involve planning, monitoring and regulation (Zimmermann and Schunk, 1989). In PISA 2003, learning strategies were measured with respect to learning of mathematics.

Seven items measuring preference for control strategies were piloted in the field trial in 2002. Five of these items were retained for the main study. All of them were inverted for scaling and positive scores indicate preferences for this learning strategy. Item wording and model parameters are shown in Table 17.18.

Table 17.17 ■ **Reliabilities for indices on self-related cognitions**

		MATHEFF	ANXMAT	SCMAT
OECD countries	Australia	0.86	0.82	0.89
	Austria	0.80	0.85	0.89
	Belgium	0.82	0.81	0.89
	Canada	0.85	0.86	0.91
	Czech Republic	0.80	0.83	0.89
	Denmark	0.83	0.85	0.90
	Finland	0.85	0.81	0.92
	France	0.78	0.75	0.89
	Germany	0.81	0.86	0.91
	Greece	0.75	0.80	0.86
	Hungary	0.82	0.81	0.81
	Iceland	0.87	0.85	0.93
	Ireland	0.81	0.83	0.89
	Italy	0.78	0.75	0.91
	Japan	0.87	0.82	0.88
	Korea	0.87	0.77	0.88
	Luxembourg	0.82	0.84	0.89
	Mexico	0.80	0.65	0.78
	Netherlands	0.83	0.81	0.90
	New Zealand	0.86	0.81	0.87
	Norway	0.84	0.85	0.90
	Poland	0.82	0.84	0.87
	Portugal	0.82	0.78	0.89
	Slovak Republic	0.83	0.81	0.87
	Spain	0.81	0.75	0.89
	Sweden	0.85	0.83	0.89
	Switzerland	0.82	0.84	0.90
	Turkey	0.85	0.82	0.88
	United Kingdom	0.86	0.84	0.88
	United States	0.86	0.87	0.89
	OECD median	**0.82**	**0.82**	**0.89**
Partner countries	Brazil	0.79	0.70	0.83
	Hong Kong-China	0.87	0.83	0.89
	Indonesia	0.74	0.70	0.75
	Latvia	0.78	0.78	0.85
	Liechtenstein	0.81	0.80	0.89
	Macao-China	0.81	0.85	0.89
	Russian Federation	0.80	0.78	0.81
	Serbia	0.79	0.82	0.83
	Thailand	0.84	0.77	0.78
	Tunisia	0.79	0.67	0.88
	Uruguay	0.82	0.75	0.88

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

Table 17.18 ■ **Item parameters for control strategies (CSTRAT)**

	There are different ways of studying mathematics: To what extent do you agree with the following statements?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST34Q01	a) When I study for a mathematics test, I try to work out what are the most important parts to learn.	-0.43	-1.68	-1.02	2.70
ST34Q03	c) When I study mathematics, I make myself check to see if I remember the work I have already done.	0.14	-2.43	-0.40	2.83
ST34Q04	d) When I study mathematics, I try to figure out which concepts I still have not understood properly.	-0.32	-1.88	-0.99	2.86
ST34Q10	j) When I cannot understand something in mathematics, I always search for more information to clarify the problem.	0.39	-2.19	-0.45	2.64
ST34Q12	l) When I study mathematics, I start by working out exactly what I need to learn.	0.21	-2.21	-0.56	2.77

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

Six items measuring preference for elaboration as a learning strategy were piloted in the field trial. Five of these items were retained for the main study. All of them are inverted for scaling, and positive scores indicate preferences for this learning strategy. Table 17.19 displays the wording of items and the parameters used for IRT scaling.

Table 17.19 ■ **Item parameters for elaboration strategies (ELAB)**

	There are different ways of studying mathematics: To what extent do you agree with the following statements?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST34Q02	b) When I am solving mathematics problems, I often think of new ways to get the answer.	-0.06	-2.58	0.14	2.44
ST34Q05	e) I think how the mathematics I have learnt can be used in everyday life.	-0.07	-2.19	-0.07	2.27
ST34Q08	h) I try to understand new concepts in mathematics by relating them to things I already know.	-0.36	-2.26	-0.43	2.69
ST34Q11	k) When I am solving a mathematics problem, I often think about how the solution might be applied to other interesting questions.	0.32	-2.59	0.20	2.39
ST34Q14	n) When learning mathematics, I try to relate the work to things I have learnt in other subjects.	0.17	-2.45	0.11	2.34

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

Five items measuring preference for memorisation/rehearsal as a learning strategy for mathematics were piloted in the field trial. Four of these items were retained for the main study. All of them were inverted for scaling, and positive scores indicate preferences for this learning strategy. Item wording and IRT model parameters are shown in Table 17.20.

Table 17.20 ■ **Item parameters for memorisation/rehearsal strategies (MEMOR)**

	There are different ways of studying mathematics: To what extent do you agree with the following statements?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST34Q06	f) I go over some problems in mathematics so often that I feel as if I could solve them in my sleep.	0.54	-1.73	0.37	1.37
ST34Q07	g) When I study for mathematics, I try to learn the answers to problems off by heart.	0.35	-1.84	-0.06	1.90
ST34Q09	i) In order to remember the method for solving a mathematics problem, I go through examples again and again.	-0.30	-1.83	-0.32	2.15
ST34Q13	m) To learn mathematics, I try to remember every step in a procedure.	-0.58	-1.91	-0.55	2.45

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

Table 17.21 ■ **Model fit and estimated latent correlations for learning strategies**

	Model fit				Latent correlations between:		
	RMSEA	RMR	CFI	NNFI	CSTRAT/ ELAB	CSTRAT/ MEMOR	ELAB/ MEMOR
Australia	0.066	0.023	0.90	0.90	0.75	0.96	0.78
Austria	0.055	0.041	0.92	0.92	0.45	0.84	0.35
Belgium	0.071	0.037	0.87	0.87	0.62	0.88	0.60
Canada	0.082	0.042	0.89	0.89	0.60	0.89	0.63
Czech Republic	0.061	0.021	0.91	0.91	0.81	0.94	0.83
Denmark	0.063	0.030	0.92	0.92	0.60	0.95	0.60
Finland	0.074	0.039	0.82	0.83	0.71	0.90	0.68
France	0.068	0.029	0.83	0.83	0.50	0.81	0.39
Germany	0.074	0.024	0.90	0.90	0.84	0.89	0.73
Greece	0.058	0.029	0.90	0.90	0.65	0.95	0.57
Hungary	0.071	0.034	0.86	0.86	0.62	0.85	0.60
Iceland	0.076	0.030	0.89	0.89	0.78	1.03	0.86
Ireland	0.093	0.046	0.85	0.85	0.81	1.00	0.76
Italy	0.057	0.040	0.93	0.93	0.49	0.94	0.50
Japan	0.066	0.022	0.92	0.92	0.89	0.91	0.86
Korea	0.056	0.026	0.92	0.92	0.56	0.96	0.74
Luxembourg	0.074	0.028	0.87	0.87	0.61	0.87	0.68
Mexico	0.077	0.040	0.87	0.87	0.80	0.97	0.86
Netherlands	0.070	0.024	0.87	0.87	0.75	0.93	0.73
New Zealand	0.067	0.028	0.85	0.85	0.62	0.94	0.48
Norway	0.089	0.037	0.88	0.88	0.74	0.78	0.62
Poland	0.066	0.030	0.92	0.92	0.71	1.03	0.75
Portugal	0.073	0.029	0.87	0.87	0.82	0.80	0.79
Slovak Republic	0.067	0.036	0.88	0.88	0.43	0.73	0.51
Spain	0.062	0.039	0.91	0.91	0.43	0.76	0.48
Sweden	0.065	0.028	0.94	0.94	0.92	0.99	0.86
Switzerland	0.071	0.029	0.80	0.81	0.63	0.64	0.18
Turkey	0.061	0.028	0.92	0.92	0.65	0.89	0.62
United Kingdom	0.063	0.030	0.93	0.93	0.75	0.89	0.82
United States	0.075	0.031	0.92	0.92	0.79	0.98	0.87
OECD	**0.054**	**0.023**	**0.94**	**0.92**	**0.66**	**0.90**	**0.67**

Note: Model estimates based on international student calibration sample (500 students per OECD country).

The CFA shows that the three-factor model has a satisfactory fit both at the international level and across

OECD countries (see Table 17.21). The fit for a multiple-group model with constrained factor loadings (RMSEA = 0.071, RMR = 0.051, CFI = 0.88, NNFI = 0.87) is only marginally different from the one with unrestricted loadings (RMSEA = 0.071, RMR = 0.030, CFI = 0.89, NNFI = 0.87). This provides some evidence of parameter invariance across OECD countries. However, in a number of countries the dimensions MEMOR and CSTRAT have estimated correlations of 1 so that in these cases the items measuring these constructs are rather one-dimensional.[7] Furthermore, scale reliabilities for MEMOR were substantially lower than for the other two indices (see Table 17.22).

Learning preferences in mathematics

Learning behaviour is also influenced by the students' preference for learning situations. Here the most salient aspects are, preference for co-operative learning for example, learning in groups, (Marsh, 1999) and preference for competitive learning, for example striving to be better than others (Owens and Barnes, 1992). Cognitive and non-cognitive benefits of co-operative goal structures have been investigated in the past. Slavin (1983) showed in a meta-analysis of studies in this field that (task-specific) co-operative learning methods per se do not affect achievement. However, co-operative learning including both individual accountability and group rewards/goals were reported to have positive effects on achievement. In PISA 2000 students that preferred either competitive or co-operative learning methods tended to perform better than other students (OECD, 2001). In PISA 2003, learning preferences were measured with respect to the learning of mathematics as the major domain.

Five items measuring preferences for competitive learning situations were piloted in the field trial. All of these items were retained for the main study. All of them are inverted for scaling and positive scores on this index indicate preferences for competitive learning situations. Item wording and parameters are shown in Table 17.23.

Table 17.22 ■ **Reliabilities for learning strategy indices**

		CSTRAT	ELAB	MEMOR
OECD countries	Australia	0.75	0.74	0.64
	Austria	0.69	0.73	0.58
	Belgium	0.72	0.74	0.48
	Canada	0.78	0.77	0.61
	Czech Republic	0.68	0.61	0.51
	Denmark	0.65	0.72	0.65
	Finland	0.68	0.74	0.70
	France	0.76	0.71	0.61
	Germany	0.72	0.73	0.68
	Greece	0.67	0.67	0.48
	Hungary	0.68	0.64	0.58
	Iceland	0.74	0.72	0.69
	Ireland	0.68	0.67	0.60
	Italy	0.73	0.74	0.49
	Japan	0.75	0.77	0.46
	Korea	0.79	0.73	0.46
	Luxembourg	0.73	0.76	0.59
	Mexico	0.73	0.74	0.65
	Netherlands	0.70	0.72	0.45
	New Zealand	0.75	0.72	0.64
	Norway	0.70	0.73	0.63
	Poland	0.67	0.68	0.52
	Portugal	0.78	0.75	0.59
	Slovak Republic	0.68	0.67	0.45
	Spain	0.75	0.74	0.57
	Sweden	0.65	0.73	0.62
	Switzerland	0.70	0.72	0.66
	Turkey	0.81	0.77	0.49
	United Kingdom	0.70	0.72	0.64
	United States	0.79	0.80	0.66
	OECD median	**0.72**	**0.73**	**0.60**
Partner countries	Brazil	0.70	0.72	0.46
	Hong Kong-China	0.76	0.80	0.58
	Indonesia	0.69	0.57	0.66
	Latvia	0.63	0.60	0.48
	Liechtenstein	0.77	0.71	0.72
	Macao-China	0.63	0.74	0.49
	Russian Federation	0.67	0.70	0.56
	Serbia	0.76	0.74	0.51
	Thailand	0.64	0.70	0.63
	Tunisia	0.75	0.70	0.59
	Uruguay	0.68	0.73	0.52

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

Table 17.23 ■ **Item parameters for preference for competitive learning situations (COMPLRN)**

	Thinking about your <mathematics> classes: To what extent do you agree with the following statements?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST37Q01	a) I would like to be the best in my class in mathematics.	-0.80	-2.81	0.40	2.40
ST37Q03	c) I try very hard in mathematics because I want to do better in the exams than the others.	-0.05	-3.31	0.29	3.02
ST37Q05	e) I make a real effort in mathematics because I want to be one of the best.	0.06	-3.15	0.32	2.83
ST37Q07	g) In mathematics I always try to do better than the other students in my class.	0.40	-3.22	0.40	2.82
ST37Q10	j) I do my best work in mathematics when I try to do better than others.	0.39	-3.03	0.29	2.74

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

Five items measuring preferences for cooperative learning situations were in the field trial. All of these items were retained for the main study. All items were inverted for scaling and positive scores on index indicate preferences for co-operative learning situations. Table 17.24 shows item wording and IRT model parameters.

Table 17.24 ■ **Item parameters for preference for co-operative learning situations (COOPLRN)**

	Thinking about your <mathematics> classes: To what extent do you agree with the following statements?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST37Q02	b) In mathematics I enjoy working with other students in groups.	-0.16	-2.28	-0.48	2.76
ST37Q04	d) When we work on a project in mathematics, I think that it is a good idea to combine the ideas of all the students in a group.	-0.36	-2.22	-0.62	2.84
ST37Q06	f) I do my best work in mathematics when I work with other students.	0.31	-2.62	-0.05	2.67
ST37Q08	h) In mathematics, I enjoy helping others to work well in a group.	0.00	-2.44	-0.46	2.90
ST37Q09	i) In mathematics I learn most when I work with other students in my class.	0.22	-2.53	-0.03	2.56

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

The CFA results in Table 17.25 show that the two-dimensional model has only a moderate to poor model fit across OECD countries. However, the model fit improves substantially (RMSEA decreases from 0.08 to 0.06) after introducing correlated error terms between items ST37Q06 and ST37Q09, which have a very similar wording; the estimated correlation between the error terms is 0.23 for the pooled international sample. The estimated latent correlation between constructs varies quite a lot between country sub-samples; this might be due to differences in learning culture and practices across educational systems.

The fit for a multiple-group model with restricted factor loadings (RMSEA = 0.0100, RMR = 0.050, CFI = 0.89, NNFI = 0.88) is not very different from the one for an unrestricted multiple-group model (RMSEA = 0.0100, RMR = 0.041, CFI = 0.91, NNFI = 0.88). However, neither model-fit indices are satisfactory. Table 17.26 shows that both indices have high reliabilities across participating countries.

Table 17.25 ■ **Model fit and estimated latent correlations for learning preferences**

	Model fit				Latent correlations between: COMPLRN/ COOPLRN
	RMSEA	RMR	CFI	NNFI	
Australia	0.090	0.035	0.94	0.92	0.20
Austria	0.092	0.049	0.91	0.91	0.27
Belgium	0.087	0.034	0.92	0.92	0.26
Canada	0.099	0.043	0.92	0.92	0.12
Czech Republic	0.077	0.028	0.93	0.93	0.08
Denmark	0.089	0.051	0.92	0.92	0.31
Finland	0.108	0.048	0.91	0.91	-0.01
France	0.101	0.051	0.88	0.88	0.23
Germany	0.081	0.048	0.93	0.93	0.29
Greece	0.094	0.035	0.91	0.91	0.45
Hungary	0.092	0.039	0.89	0.89	0.17
Iceland	0.133	0.076	0.86	0.86	0.25
Ireland	0.086	0.033	0.92	0.92	0.18
Italy	0.094	0.038	0.92	0.92	0.24
Japan	0.092	0.057	0.93	0.93	0.44
Korea	0.093	0.030	0.93	0.93	0.84
Luxembourg	0.102	0.048	0.92	0.92	0.29
Mexico	0.092	0.027	0.92	0.92	0.70
Netherlands	0.063	0.018	0.95	0.95	0.24
New Zealand	0.095	0.037	0.92	0.92	0.18
Norway	0.116	0.086	0.87	0.87	0.15
Poland	0.093	0.032	0.90	0.90	0.35
Portugal	0.082	0.027	0.94	0.94	0.57
Slovak Republic	0.075	0.030	0.94	0.94	-0.02
Spain	0.065	0.025	0.96	0.96	0.50
Sweden	0.076	0.037	0.95	0.95	0.26
Switzerland	0.078	0.044	0.93	0.93	0.09
Turkey	0.067	0.035	0.96	0.96	0.62
United Kingdom	0.110	0.044	0.89	0.89	0.33
United States	0.104	0.041	0.92	0.92	0.48
OECD	**0.080**	**0.032**	**0.94**	**0.93**	**0.35**

Note: Model estimates based on international student calibration sample (500 students per OECD country).

Table 17.26 ■ **Reliabilities for indices of learning preferences**

		COMPLRN	COOPLRN
OECD countries	Australia	0.86	0.80
	Austria	0.82	0.75
	Belgium	0.83	0.74
	Canada	0.86	0.81
	Czech Republic	0.84	0.74
	Denmark	0.87	0.74
	Finland	0.86	0.74
	France	0.82	0.75
	Germany	0.84	0.77
	Greece	0.81	0.79
	Hungary	0.82	0.69
	Iceland	0.83	0.80
	Ireland	0.82	0.77
	Italy	0.85	0.77
	Japan	0.87	0.80
	Korea	0.83	0.79
	Luxembourg	0.85	0.83
	Mexico	0.82	0.77
	Netherlands	0.83	0.71
	New Zealand	0.85	0.80
	Norway	0.84	0.76
	Poland	0.80	0.75
	Portugal	0.84	0.76
	Slovak Republic	0.82	0.75
	Spain	0.85	0.79
	Sweden	0.84	0.78
	Switzerland	0.83	0.76
	Turkey	0.85	0.82
	United Kingdom	0.84	0.78
	United States	0.86	0.86
	OECD median	0.84	0.77
Partner countries	**Brazil**	**0.80**	**0.74**
	Hong Kong-China	0.81	0.80
	Indonesia	0.77	0.59
	Latvia	0.70	0.70
	Liechtenstein	0.82	0.76
	Macao-China	0.77	0.74
	Russian Federation	0.81	0.72
	Serbia	0.84	0.83
	Thailand	0.78	0.71
	Tunisia	0.83	0.77
	Uruguay	0.81	0.76

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

Classroom climate

The learning climate at the classroom level is an important factor in explaining student performance. Many classroom observation studies have provided evidence of the influence of general school environment factors on instructional quality in the classroom, as well as the effects of teacher behaviour and classroom practice on the overall performance of schools (see for example Schaffer *et al.*, 1994).

In PISA 2000, five items regarding perceived teacher support in test language classes were administered. In PISA 2003, similar items regarding teacher support in mathematics classes were included. All items were inverted so that positive scores on this index indicate perceptions of higher levels of teacher support. Table 17.27 shows the item wording and the IRT model parameters.

Table 17.27 ■ **Item parameters for teacher support (TEACHSUP)**

	How often do these things happen in your <mathematics> lessons?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST38Q01	a) The teacher shows an interest in every student's learning.	0.26	-2.00	0.09	1.91
ST38Q03	c) The teacher gives extra help when students need it.	-0.14	-1.73	0.04	1.69
ST38Q05	e) The teacher helps students with their learning.	-0.49	-1.67	-0.01	1.67
ST38Q07	g) The teacher continues teaching until the students understand.	0.08	-1.61	0.02	1.59
ST38Q10	j) The teacher gives students an opportunity to express opinions.	0.30	-1.65	-0.09	1.74

Note: Categories were "every lesson", "most lessons", "some lessons" and "never or hardly ever". All items were inverted for scaling.

Disciplinary climate in language classes in PISA 2000 was measured using five items. Similar items (with minor modifications) asking about the disciplinary climate in mathematics classes were used again in PISA 2003. The items were not inverted for scaling so that positive scores on the index indicate perceptions of a positive disciplinary climate. Item wording and item parameters are illustrated in Table 17.28.

Table 17.28 ■ **Item parameters for disciplinary climate (DISCLIM)**

	How often do these things happen in your <mathematics> lessons?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST38Q02	b) Students don't listen to what the teacher says.	0.12	-2.09	-0.67	2.76
ST38Q06	f) There is noise and disorder.	0.38	-1.28	-0.65	1.93
ST38Q08	h) The teacher has to wait a long time for students to <quieten down>.	0.07	-1.31	-0.54	1.84
ST38Q09	i) Students cannot work well.	-0.41	-1.61	-0.61	2.22
ST38Q11	k) Students don't start working for a long time after the lesson begins.	-0.16	-1.39	-0.44	1.84

Note: Categories were "every lesson", "most lessons", "some lessons" and "never or hardly ever".

Table 17.29 ■ **Model fit and estimated latent correlations for learning preferences**

	Model fit				Latent correlations between:
	RMSEA	RMR	CFI	NNFI	TEACHSUP/ DISCLIM
Australia	0.078	0.031	0.96	0.96	0.32
Austria	0.048	0.040	0.98	0.98	0.18
Belgium	0.053	0.030	0.98	0.98	0.12
Canada	0.062	0.036	0.96	0.97	0.30
Czech Republic	0.059	0.037	0.97	0.97	0.24
Denmark	0.061	0.028	0.96	0.96	0.35
Finland	0.070	0.032	0.96	0.96	0.31
France	0.058	0.043	0.97	0.97	0.20
Germany	0.029	0.028	0.99	0.99	0.23
Greece	0.044	0.034	0.97	0.97	0.23
Hungary	0.062	0.037	0.97	0.97	0.23
Iceland	0.044	0.021	0.98	0.98	0.18
Ireland	0.063	0.047	0.97	0.97	0.35
Italy	0.036	0.037	0.99	0.99	0.26
Japan	0.049	0.036	0.97	0.97	0.23
Korea[a]	N/A	N/A	N/A	N/A	N/A
Luxembourg	0.042	0.047	0.98	0.98	0.10
Mexico	0.065	0.044	0.95	0.95	0.22
Netherlands	0.033	0.028	0.99	0.99	0.33
New Zealand	0.058	0.026	0.98	0.98	0.35
Norway	0.053	0.031	0.97	0.97	0.24
Poland	0.067	0.031	0.96	0.96	0.18
Portugal	0.047	0.029	0.98	0.98	0.15
Slovak Republic	0.032	0.027	0.99	0.99	0.07
Spain	0.040	0.036	0.99	0.99	0.09
Sweden	0.055	0.032	0.97	0.97	0.29
Switzerland	0.017	0.022	1.00	1.00	0.08
Turkey	0.056	0.040	0.96	0.96	0.05
United Kingdom	0.056	0.035	0.98	0.98	0.42
United States	0.085	0.034	0.93	0.94	0.25
OECD	**0.039**	**0.019**	**0.99**	**0.98**	**0.20**

Note: Model estimates based on international student calibration sample (500 students per OECD country).
a. There are no model estimates available for Korea due to the incorrect translation of item 38f.

Table 17.29 shows the CFA results for the OECD countries: the fit for the two-dimensional model is highly satisfactory at both the national and international level. There is a considerable variation in the estimated latent correlation between both dimensions. This is plausible in view of differences in classroom practices and learning culture across OECD countries.

The fit for the multiple-group model with unrestricted factor loadings (RMSEA = 0.055, RMR = 0.034, CFI = 0.97, NNFI = 0.96) is only marginally superior to the model with constrained factor loadings (RMSEA = 0.057, RMR = 0.042, CFI = 0.96, NNFI = 0.96). This supports the assumption of parameter invariance across OECD countries. Both indices have highly satisfactory reliabilities across all PISA countries (Table 17.30).

Table 17.30 ■ **Reliabilities for indices of classroom climate**

		TEACHSUP	DISCLIM
OECD countries	Australia	0.87	0.87
	Austria	0.83	0.85
	Belgium	0.85	0.84
	Canada	0.85	0.83
	Czech Republic	0.80	0.84
	Denmark	0.80	0.84
	Finland	0.83	0.85
	France	0.82	0.84
	Germany	0.85	0.84
	Greece	0.79	0.72
	Hungary	0.85	0.84
	Iceland	0.81	0.85
	Ireland	0.85	0.86
	Italy	0.86	0.85
	Japan	0.78	0.79
	Korea[a]	0.75	0.73
	Luxembourg	0.87	0.81
	Mexico	0.83	0.75
	Netherlands	0.78	0.83
	New Zealand	0.86	0.86
	Norway	0.81	0.82
	Poland	0.84	0.84
	Portugal	0.87	0.80
	Slovak Republic	0.79	0.80
	Spain	0.85	0.83
	Sweden	0.83	0.82
	Switzerland	0.80	0.82
	Turkey	0.82	0.79
	United Kingdom	0.87	0.88
	United States	0.86	0.83
	OECD median	**0.83**	**0.83**
Partner countries	Brazil	0.82	0.70
	Hong Kong-China	0.83	0.88
	Indonesia	0.60	0.74
	Latvia	0.76	0.83
	Liechtenstein	0.81	0.85
	Macao-China	0.78	0.83
	Russian Federation	0.75	0.83
	Serbia	0.85	0.79
	Thailand	0.81	0.80
	Tunisia	0.79	0.68
	Uruguay	0.84	0.82

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.
a. The reliability coefficient for DISCLIM for Korea is based on four items only due to a translation error for item 38f.

ICT familiarity scale indices

The scale indices described in this section are only available for those countries which chose to administer the international option of an ICT familiarity questionnaire. The ICT familiarity questionnaire included items regarding ICT use, attitudes towards computers, and self-confidence with ICT tasks.

Computer use and attitudes

In the field trial, six items measuring the frequency of different types of ICT use were related to a common dimension called ICT Internet and Entertainment use and were retained for the main study. All items were inverted for scaling and positive scores on this index indicate high frequencies of ICT use. Table 17.31 contains item wording and IRT model parameters.

Table 17.31 ■ **Item parameters for Internet/entertainment use (INTUSE)**

	How often do you use:	Parameter estimates				
		Delta	Tau(1)	Tau(2)	Tau(3)	Tau(4)
IC05Q01	a) The Internet to look up information about people, things, or ideas?	-0.32	-0.66	-0.73	-0.05	1.43
IC05Q02	b) Games on a computer?	-0.25	-0.59	-0.24	-0.10	0.93
IC05Q04	d) The Internet to collaborate with a group or team?	0.45	-0.22	-0.62	-0.07	0.91
IC05Q06	f) The Internet to download software?	0.29	0.03	-0.50	-0.21	0.68
IC05Q10	j) The Internet to download music?	0.03	0.39	-0.50	-0.29	0.40
IC05Q12	l) A computer for electronic communication (*e.g.* e-mail or "chat rooms")?	-0.20	0.25	-0.45	-0.32	0.52

Note: Categories were "almost every day", "a few times each week", "between once a week and once a month", "less than once a month" and "never". All items were inverted for scaling.

In the field trial, six items were related to a dimension called ICT program/software use and were retained for the main study. All items were inverted for scaling, and positive scores on this index indicate high frequencies of ICT use. Item wording and IRT model parameters are shown in Table 17.32.

Table 17.32 ■ **Item parameters for programme/software use (INTPRG)**

	How often do you use:	Parameter estimates				
		Delta	Tau(1)	Tau(2)	Tau(3)	Tau(4)
IC05Q03	c) Word processing (*e.g.* Microsoft® Word® or Word Perfect®)?	-0.73	-0.87	-0.90	0.02	1.75
IC05Q05	e) Spreadsheets (*e.g.* Lotus 1-2-3® or Microsoft® Excel®)?	0.23	-0.82	-0.58	0.03	1.38
IC05Q07	g) Drawing, painting or graphics programs on a computer?	-0.17	-0.86	-0.37	0.11	1.12
IC05Q08	h) Educational software such as mathematics programs?	0.60	-0.52	-0.50	0.04	0.98
IC05Q09	i) The computer to help you learn school material?	-0.08	-0.58	-0.68	-0.07	1.32
IC05Q11	k) The computer for programming?	0.15	-0.11	-0.42	-0.06	0.60

Note: Categories were "almost every day", "a few times each week", "between once a week and once a month", "less than once a month" and "never". All items were inverted for scaling.

Four items measuring attitudes towards computers were used in PISA 2003. All items were inverted for scaling so that positive scores on this index indicate positive attitudes towards computers.[8] Item wording and IRT model parameters are shown in Table 17.33.

Table 17.33 ■ **Item parameters for attitudes towards computers (ATTCOMP)**

	How much do you disagree or agree with the following statements about you and computers?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
IC07Q01	a) It is very important to me to work with a computer.	-0.17	-2.33	-0.06	2.39
IC07Q02	b) To play or work with a computer is really fun.	-0.61	-1.99	-0.64	2.63
IC07Q03	c) I use a computer because I am very interested.	0.19	-2.50	0.17	2.32
IC07Q04	d) I lose track of time, when I am working with the computer.	0.58	-2.31	0.26	2.05

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling.

Table 17.34 ■ **Model fit and estimated latent correlations for ICT use and attitudes**

	Model fit				Latent correlations between:		
	RMSEA	RMR	CFI	NNFI	INTUSE/ PRGUSE	INTUSE/ ATTCOMP	PRGUSE/ ATTCOMP
Australia	0.104	0.094	0.83	0.80	0.73	0.66	0.63
Austria	0.107	0.137	0.81	0.77	0.55	0.46	0.47
Belgium	0.088	0.109	0.88	0.86	0.59	0.63	0.48
Canada	0.089	0.102	0.87	0.85	0.66	0.67	0.37
Czech Republic	0.105	0.119	0.82	0.78	0.59	0.50	0.46
Denmark	0.111	0.116	0.83	0.80	0.79	0.76	0.63
Finland	0.099	0.088	0.84	0.81	0.73	0.71	0.53
Germany	0.113	0.139	0.80	0.76	0.61	0.51	0.57
Greece	0.082	0.095	0.90	0.88	0.71	0.51	0.44
Hungary	0.085	0.095	0.86	0.83	0.68	0.38	0.45
Iceland	0.095	0.096	0.86	0.84	0.79	0.67	0.47
Ireland	0.082	0.086	0.90	0.88	0.66	0.58	0.51
Italy	0.074	0.118	0.90	0.88	0.56	0.49	0.51
Japan	0.078	0.070	0.91	0.89	0.73	0.55	0.37
Korea	0.095	0.100	0.80	0.76	0.52	0.49	0.19
Mexico	0.081	0.105	0.89	0.87	0.71	0.50	0.44
New Zealand	0.086	0.093	0.88	0.86	0.61	0.59	0.48
Poland	0.108	0.124	0.86	0.84	0.71	0.43	0.33
Portugal	0.094	0.102	0.88	0.86	0.70	0.49	0.45
Slovak Republic	0.100	0.142	0.85	0.82	0.63	0.41	0.50
Sweden	0.090	0.095	0.87	0.85	0.71	0.72	0.56
Switzerland	0.093	0.108	0.87	0.85	0.65	0.59	0.56
Turkey	0.078	0.094	0.91	0.89	0.75	0.47	0.47
United States	0.082	0.096	0.87	0.84	0.69	0.57	0.42
OECD	**0.088**	**0.099**	**0.88**	**0.85**	**0.65**	**0.55**	**0.48**

Note: Model estimates based on international student calibration sample (500 students per OECD country).

The CFA results in Table 17.34 show only moderate model fit for the three-dimensional model across sub-samples, in some countries the fit for this model is unsatisfactory. All three constructs are highly correlated with each other; the two constructs related to use of ICT have the highest correlation. The model fit for the multiple-group model with unconstrained parameters (RMSEA = 0.098, RMR = 0.096, CFI = 0.86, NNFI = 0.84) is only marginally different from the one for the model with constrained factor loadings (RMSEA = 0.098, RMR = 0.110, CFI = 0.85, NNFI = 0.84); this provides some support for the assumption of parameter invariance. However, it should be noted that both models have rather poor fit indices across OECD countries.

Table 17.35 shows that the scale reliabilities for all three indices are satisfactory across all the countries that administered the ICT familiarity questionnaire.

Table 17.35 ■ **Reliabilities for computer use and attitudes**

		INTUSE	PRGUSE	ATTCOMP
OECD countries	**Australia**	0.80	0.78	0.80
	Austria	0.79	0.76	0.82
	Belgium	0.83	0.77	0.83
	Canada	0.79	0.79	0.82
	Czech Republic	0.83	0.78	0.78
	Denmark	0.77	0.79	0.84
	Finland	0.79	0.78	0.83
	Germany	0.81	0.77	0.82
	Greece	0.85	0.82	0.81
	Hungary	0.77	0.74	0.82
	Iceland	0.78	0.81	0.81
	Ireland	0.82	0.80	0.81
	Italy	0.82	0.77	0.78
	Japan	0.80	0.70	0.89
	Korea	0.67	0.77	0.81
	Mexico	0.87	0.83	0.72
	New Zealand	0.81	0.80	0.80
	Poland	0.86	0.87	0.85
	Portugal	0.85	0.79	0.82
	Slovak Republic	0.82	0.82	0.78
	Sweden	0.78	0.76	0.82
	Switzerland	0.81	0.79	0.85
	Turkey	0.86	0.85	0.80
	United Kingdom	0.82	0.79	0.81
	United States	0.79	0.79	0.77
	OECD median	**0.81**	**0.79**	**0.81**
Partner countries	**Latvia**	0.84	0.81	0.78
	Liechtenstein	0.81	0.79	0.84
	Russian Federation	0.88	0.81	0.85
	Serbia	0.88	0.82	0.86
	Thailand	0.87	0.80	0.72
	Tunisia	0.86	0.88	0.75
	Uruguay	0.85	0.85	0.70

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

ICT self-confidence

In the field trial, 14 items on self-confidence with ICT tasks were included to measure the dimension confidence in routine tasks (ROUTCONF) and 11 of these items were retained for the main study. All items were inverted for scaling so that positive WLEs indicate high self-confidence. Table 17.36 contains the item wording and the IRT parameters used for scaling.

Table 17.36 ■ **Item parameters for confidence in routine tasks (ROUTCONF)**

	How well can you do each of these tasks on a computer?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
IC06Q01	a) Start a computer game.	-0.31	-0.90	0.00	0.90
IC06Q03	c) Open a file.	-0.42	-0.43	-0.09	0.52
IC06Q04	d) Create/edit a document.	0.37	-0.75	-0.08	0.83
IC06Q05	e) Scroll a document up and down a screen.	0.23	-0.21	0.04	0.17
IC06Q07	g) Copy a file from a floppy disk.	0.61	-1.15	0.14	1.01
IC06Q08	h) Save a computer document or file.	-0.17	-0.64	0.16	0.47
IC06Q09	i) Print a computer document or file.	-0.13	-0.87	0.28	0.59
IC06Q10	j) Delete a computer document or file.	-0.17	-0.63	0.13	0.49
IC06Q11	k) Moves files form one place to another on a computer.	0.49	-1.09	-0.05	1.14
IC06Q18	r) Play computer games.	-0.43	-0.51	0.07	0.44
IC06Q21	u) Draw pictures using a mouse.	-0.08	-0.83	0.20	0.63

Note: Categories were "I can do this very well by myself", "I can do this with help from someone", "I know what this is but I cannot do it" and "I don't know what this means". All items were inverted for scaling.

In the field trial, five items on self-confidence with ICT tasks were related to the dimension confidence in Internet tasks (INTCONF), all of them were retained for the main study. All items were inverted for scaling and positive scores on this index indicate high self-confidence. Item wording and IRT model parameters are shown in Table 17.37.

Table 17.37 ■ **Item parameters for confidence in Internet tasks (INTCONF)**

	How well can you do each of these tasks on a computer?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
IC06Q12	l) Get on to the Internet.	-1.00	-1.00	0.53	0.47
IC06Q13	m) Copy or download files from the Internet.	0.01	-1.70	0.14	1.56
IC06Q14	n) Attach a file to an email message.	0.88	-1.65	0.03	1.62
IC06Q19	s) Download music from the Internet.	0.28	-1.96	0.30	1.66
IC06Q22	v) Write and send emails.	-0.17	-1.21	0.30	0.91

Note: Categories were "I can do this very well by myself", "I can do this with help from someone", "I know what this is but I cannot do it" and "I don't know what this means". All items were inverted for scaling.

Six items were included to measure confidence in ICT high level tasks (HIGHCONF). All items were inverted for scaling, and positive scores on index indicate high self-confidence. Table 17.38 shows the item wording and the IRT model parameters used for scaling.

Table 17.38 ■ **Item parameters for confidence in ICT high level tasks (HIGHCONF)**

	How well can you do each of these tasks on a computer?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
IC06Q02	b) Use software to find and get rid of computer viruses.	-0.05	-1.64	0.45	1.18
IC06Q06	f) Use a database to produce a list of addresses.	-0.44	-0.66	-0.34	1.00
IC06Q15	o) Create a computer program (e.g. in Logo, Pascal, Basic).	0.70	-1.61	0.07	1.55
IC06Q16	p) Use a spreadsheet to plot a graph.	-0.13	-0.95	-0.10	1.05
IC06Q17	q) Create a presentation (*e.g.* using <Microsoft® PowerPoint®>).	-0.12	-0.66	-0.11	0.78
IC06Q20	t) Create a multi-media presentation (with sound, pictures, video).	-0.04	-1.56	0.13	1.43
IC06Q23	h) Construct a Web page.	0.08	-1.94	0.16	1.78

Note: Categories were "I can do this very well by myself", "I can do this with help from someone", "I know what this is but I cannot do it" and "I don't know what this means". All items were inverted for scaling.

Table 17.39 ■ **Model fit and estimated latent correlations for ICT confidence items**

	Model fit				Latent correlations between:		
	RMSEA	INTCONF	HIGHCONF	HIGHCONF	ROUTCONF/ INTCONF	ROUTCONF/ HIGHCONF	INTCONF/ HIGHCONF
Australia	0.108	0.022	0.73	0.70	0.59	0.57	0.74
Austria	0.109	0.031	0.70	0.67	0.63	0.40	0.77
Belgium	0.092	0.032	0.83	0.81	0.64	0.46	0.54
Canada	0.129	0.038	0.64	0.61	1.00	0.26	0.30
Czech Republic	0.094	0.035	0.72	0.69	0.56	0.61	0.79
Denmark	0.108	0.039	0.75	0.72	0.61	0.61	0.77
Finland	0.117	0.058	0.73	0.71	0.75	0.64	0.79
Germany	0.122	0.048	0.70	0.67	0.63	0.55	0.73
Greece	0.086	0.048	0.86	0.85	0.79	0.63	0.76
Hungary	0.100	0.051	0.81	0.79	0.68	0.63	0.86
Iceland	0.122	0.052	0.76	0.74	0.85	0.44	0.58
Ireland	0.103	0.071	0.80	0.78	0.66	0.59	0.80
Italy	0.092	0.065	0.83	0.81	0.58	0.52	0.70
Japan	0.097	0.066	0.86	0.84	0.75	0.66	0.78
Korea	0.121	0.039	0.66	0.62	0.94	0.37	0.25
Mexico	0.098	0.047	0.86	0.84	0.71	0.73	0.74
New Zealand	0.112	0.036	0.75	0.72	0.83	0.53	0.76
Poland	0.105	0.043	0.81	0.79	0.72	0.61	0.85
Portugal	0.101	0.042	0.87	0.85	0.64	0.62	0.88
Slovak Republic	0.100	0.083	0.84	0.83	0.62	0.64	0.84
Sweden	0.095	0.027	0.77	0.75	0.74	0.55	0.71
Switzerland	0.101	0.047	0.80	0.78	0.65	0.56	0.69
Turkey	0.105	0.078	0.83	0.81	0.70	0.71	0.80
United States	0.104	0.031	0.77	0.74	0.66	0.41	0.72
OECD	**0.081**	**0.039**	**0.87**	**0.87**	**0.71**	**0.60**	**0.74**

Note: Model estimates based on international student calibration sample (500 students per OECD country).

The CFA results in Table 17.39 show that the fit for the three-dimensional model was not entirely satisfactory across OECD countries. All three constructs are highly correlated with each other. In Canada ROUTCONF and INTCONF are perfectly correlated, which indicates that in this country the two dimensions cannot be distinguished. It should be noted that responses to the items measuring routine tasks were highly skewed with only few students expressing a lack of confidence with these tasks.

The fit for a multiple-group model with unconstrained loadings (RMSEA = 0.110, RMR = 0.029, CFI = 0.80, NNFI = 0.78) is only slightly better than the one for the model with restricted parameters (RMSEA = 0.110, RMR = 0.033, CFI = 0.77, NNFI = 0.76).

Table 17.40 shows that the internal consistency of all three indices is high across participating countries.

Table 17.40 ■ **Reliabilities for computer use and attitudes**

		ROUTCONF	INTCONF	HIGHCONF
OECD countries	Australia	0.88	0.78	0.83
	Austria	0.87	0.77	0.82
	Belgium	0.90	0.85	0.83
	Canada	0.82	0.71	0.83
	Czech Republic	0.87	0.81	0.85
	Denmark	0.82	0.71	0.85
	Finland	0.88	0.73	0.87
	Germany	0.87	0.81	0.84
	Greece	0.89	0.84	0.86
	Hungary	0.90	0.84	0.84
	Iceland	0.89	0.73	0.84
	Ireland	0.87	0.82	0.85
	Italy	0.88	0.87	0.82
	Japan	0.93	0.88	0.87
	Korea	0.81	0.72	0.81
	Mexico	0.94	0.91	0.87
	New Zealand	0.87	0.79	0.83
	Poland	0.92	0.85	0.85
	Portugal	0.93	0.88	0.86
	Slovak Republic	0.92	0.88	0.86
	Sweden	0.85	0.69	0.85
	Switzerland	0.87	0.80	0.85
	Turkey	0.93	0.88	0.86
	United Kingdom	0.88	0.77	0.83
	United States	0.89	0.77	0.83
	OECD median	**0.88**	**0.81**	**0.85**
Partner countries	Latvia	0.91	0.86	0.84
	Liechtenstein	0.88	0.71	0.85
	Russian Federation	0.94	0.93	0.88
	Serbia	0.92	0.92	0.87
	Thailand	0.91	0.90	0.87
	Tunisia	0.95	0.88	0.87
	Uruguay	0.93	0.91	0.85

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

School-level variation of student scale indices

One important aspect in the analysis of student-level indices is the extent to which there is variation in student scores across schools within participating countries. This is of particular interest with regard to indices measuring school-related attitudes or behaviour, as it provides evidence to what extent these student-level measures are influenced by school characteristics.

In order to assess the proportion of between-school variance, two-level random effects models for school intercepts of student-level indices were estimated within each country (Bryk and Raudenbush, 1992); this provides variance estimates for student and school level. Table 17.41 shows the median, maximum and minimum percentages of between-school variance. For most of the student-level indices, the proportion of between-school variance is below 10 per cent. However, for a number of indices the maximum variance between schools is more than 25 per cent.

Notably, home background indices like ESCS, COMPHOME or CULTPOSS have the highest intra-class correlations. Indices related to learning strategies (ELAB, CSTRAT, MEMOR) or preferences (COMPLRN, COOPLRN) and those attitudes toward school (BELONG, ATSCHL), however, have rather low between-school variance across countries. Among the school-related indices, mathematics self-efficacy (MATHEFF) and those related to classroom climate have the largest proportions of between-school variance.

Table 17.41 ■ **Median, minimum and maximum percentages of between-school variance for student-level indices across countries**

Index	OECD countries			Partner countries		
	Median	Minimum	Maximum	Median	Minimum	Maximum
COMPHOME	10.2	1.8	29.5	19.1	4.0	37.2
HEDRES	6.7	1.4	23.8	10.5	5.4	23.5
CULTPOSS	11.3	3.5	26.7	10.7	6.9	16.2
HOMEPOS	17.5	4.0	37.0	19.0	11.0	33.0
ATSCHL	4.1	1.1	7.6	4.5	2.7	7.2
STUREL	7.3	4.5	13.7	9.0	2.8	15.0
BELONG	3.0	0.0	6.7	3.2	0.3	9.8
INTMAT	4.2	1.9	10.5	5.7	0.0	12.5
INSTMOT	3.7	1.1	17.0	4.6	0.6	10.0
MATHEFF	9.0	2.7	26.0	8.5	6.1	16.3
ANXMAT	3.2	1.5	7.1	4.9	0.8	6.9
SCMAT	3.5	1.0	10.3	4.5	0.0	10.9
CSTRAT	2.9	0.9	13.8	2.6	1.7	5.4
ELAB	3.4	0.5	7.8	4.6	1.2	15.2
MEMOR	2.7	0.3	6.2	3.5	1.4	11.8
COMPLRN	4.3	1.2	13.6	5.3	0.0	9.4
COOPLRN	3.0	0.8	8.4	3.0	2.0	7.6
TEACHSUP	7.4	4.0	16.6	7.3	1.5	13.4
DISCLIM	9.9	4.1	27.0	9.9	5.8	18.8
INTUSE	4.9	1.4	13.9	9.2	1.4	34.4
PRGUSE	4.7	1.7	14.4	6.0	3.0	20.4
ROUTCONF	8.3	1.2	28.4	15.4	1.6	28.0
INTCONF	7.2	0.9	27.7	17.3	0.6	43.7
HIGHCONF	3.8	1.7	16.5	7.4	1.2	25.5
ATTCOMP	2.8	0.0	6.6	4.4	1.1	8.1

School questionnaire scale indices

Quality of school resources

The index of quality of schools' physical infrastructure (SCMATBUI) is derived from three items measuring the school principal's perceptions of potential factors hindering instruction at school. The indices were used in PISA 2000 but the question format had been modified. All items were inverted for scaling so that positive scores indicate positive evaluations of this aspect. Item wording and IRT model parameters are shown in Table 17.42.

Table 17.42 ■ **Item parameters for quality of the school's physical infrastructure (SCMATBUI)**

	Is your school's capacity to provide instruction hindered by a shortage or inadequacy of any of the following?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
SC08Q11	k) School buildings and grounds	0.32	-1.61	0.26	1.36
SC08Q12	l) Heating/cooling and lighting systems	-0.45	-1.61	0.12	1.49
SC08Q13	m) Instructional space (*e.g.* classrooms)	0.14	-1.86	0.32	1.54

Note: Categories were "not at all", "very little", "to some extent" and "a lot". All items were inverted for scaling.

The index of quality of school's educational resources (SCMATEDU) is derived from seven items measuring the school principal's perceptions of potential factors hindering instruction at school. All items were inverted for scaling so that positive WLEs indicate positive evaluations of this aspect. Table 17.43 illustrates the wording of the items and the IRT model parameters.

Table 17.43 ■ **Item parameters for quality of the school's educational resources (SCMATEDU)**

	Is your school's capacity to provide instruction hindered by a shortage or inadequacy of any of the following?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
SC08Q09	i) Instructional materials (*e.g.* textbooks)	-0.37	-1.71	0.23	1.48
SC08Q15	o) Computers for instruction	0.22	-1.83	0.38	1.44
SC08Q16	p) Computer software for instruction	0.50	-2.01	0.27	1.75
SC08Q17	q) Calculators for instruction	-1.08	-1.48	0.14	1.34
SC08Q18	r) Library materials	0.06	-1.85	0.16	1.70
SC08Q19	s) Audio-visual resources	0.17	-1.98	0.15	1.83
SC08Q20	t) Science laboratory equipment and materials	0.49	-1.79	0.10	1.70

Note: Categories were "not at all", "very little", "to some extent" and "a lot". All items were inverted for scaling.

The index on teacher shortage (TCSHORT) is derived from four items measuring the school principal's perceptions of potential factors hindering instruction at school. Items were not inverted for scaling so that positive WLEs indicate school principal's reports of higher levels of teacher shortage at the tested school. Table 17.44 shows the item wording and the IRT model parameters.

Table 17.44 ■ **Item parameters for teacher shortage (TCSHORT)**

	Is your school's capacity to provide instruction hindered by a shortage or inadequacy of any of the following?	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
SC08Q01	a) Availability of qualified mathematics teachers	-0.37	-1.71	0.23	1.48
SC08Q02	b) Availability of qualified science teachers	0.22	-1.83	0.38	1.44
SC08Q03	c) Availability of qualified <test language> teachers	0.50	-2.01	0.27	1.75
SC08Q05	e) Availability of qualified foreign language teachers	-1.08	-1.48	0.14	1.34
SC08Q06	f) Availability of experienced teachers	0.06	-1.85	0.16	1.70

Note: Categories were "not at all", "very little", "to some extent" and "a lot".

The CFA of a three-factor model show a satisfactory model fit (RMSEA = 0.078, RMR = 0.043, CFI = 0.93, NNFI = 0.92) for the international school calibration sample. The estimated correlations are -0.39 for SCMATBUI with TCSHORT, -0.46 for SCMATEDU with TCSHORT and 0.67 for SCMATEDU with SCMATBUI. The scale reliabilities for all three indices are satisfactory across participating countries (see Table 17.45).

Table 17.45 ■ **Reliabilities for indices of school resource quality**

		TCSHORT	SCMATBUI	SCMATEDU
OECD countries	Australia	0.78	0.75	0.87
	Austria	0.88	0.85	0.83
	Belgium	0.87	0.84	0.87
	Canada	0.75	0.81	0.90
	Czech Republic	0.52	0.79	0.71
	Denmark	0.74	0.82	0.85
	Finland	0.70	0.90	0.83
	France	N/A	N/A	N/A
	Germany	0.75	0.86	0.83
	Greece	0.98	0.90	0.80
	Hungary	0.62	0.68	0.80
	Iceland	0.84	0.80	0.85
	Ireland	0.80	0.85	0.83
	Italy	0.94	0.83	0.90
	Japan	0.95	0.85	0.93
	Korea	0.79	0.81	0.86
	Luxembourg	0.89	0.70	0.67
	Mexico	0.92	0.81	0.89
	Netherlands	0.79	0.82	0.81
	New Zealand	0.73	0.70	0.86
	Norway	0.69	0.81	0.73
	Poland	0.87	0.66	0.82
	Portugal	0.85	0.81	0.87
	Slovak Republic	0.71	0.59	0.79
	Spain	0.95	0.81	0.88
	Sweden	0.87	0.70	0.85
	Switzerland	0.87	0.77	0.82
	Turkey	0.83	0.87	0.88
	United Kingdom	0.88	0.83	0.88
	United States	0.76	0.72	0.84
	OECD median	**0.83**	**0.81**	**0.85**
Partner countries	Brazil	0.91	0.74	0.90
	Hong Kong-China	0.84	0.79	0.92
	Indonesia	0.86	0.61	0.89
	Latvia	0.56	0.72	0.83
	Liechtenstein	0.76	0.67	0.88
	Macao-China	0.90	0.79	0.88
	Russian Federation	0.80	0.87	0.80
	Serbia	0.53	0.53	0.71
	Thailand	0.75	0.87	0.90
	Tunisia	0.74	0.62	0.82
	Uruguay	0.88	0.75	0.86

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

Teacher consensus, teacher and student morale

The index of school principals' perception of teacher morale and commitment (TCMORALE) is derived from four items measuring the school principal's perceptions of teachers at a school. All items were inverted for scaling and the categories "disagree" and "strongly disagree" were collapsed into one category in view of very few responses in the these categories. Positive scores on this index indicate principals' reports of higher levels of teacher morale and commitment. Table 17.46 illustrates item wording and IRT model parameters.

Table 17.46 ■ **Item parameters for teacher morale (TCMORALE)**

	Think about the teachers in your school. How much do you agree with the following statements?	Parameter estimates		
		Delta	Tau(1)	Tau(2)
SC24Q01	a) The morale of teachers in this school is high.	0.36	-2.50	2.50
SC24Q02	b) Teachers work with enthusiasm.	0.31	-2.90	2.90
SC24Q03	c) Teachers take pride in this school.	0.06	-2.75	2.75
SC24Q04	d) Teachers value academic achievement.	-0.73	-2.76	2.76

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling and the categories "disagree" and "strongly disagree" were recoded into one category.

Table 17.47 ■ **Item parameters for student morale (STMORALE)**

	Think about the students in your school. How much do you agree with the following statements?	Parameter estimates		
		Delta	Tau(1)	Tau(2)
SC11Q01	a) Students enjoy being in school.	-1.35	-3.15	3.15
SC11Q02	b) Students work with enthusiasm.	1.02	-2.93	2.93
SC11Q03	c) Students take pride in this school.	-0.34	-2.80	2.80
SC11Q04	d) Students value academic achievement.	-0.08	-2.62	2.62
SC11Q05	e) Students are cooperative and respectful. .	-0.50	-3.11	3.11
SC11Q06	f) Students value the education they can receive in this school.	-0.25	-2.98	2.98
SC11Q07	g) Students do their best to learn as much as possible.	1.50	-2.58	2.58

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". All items were inverted for scaling and the categories "disagree" and "strongly disagree" were recoded into one category.

The index of school principals' perceptions of student morale and commitment (STMORALE) is derived from seven items measuring the school principal's perceptions of students at a school. All items were inverted for scaling, and the categories "disagree" and "strongly disagree" were collapsed into one category in view of very few responses in these categories. Positive scores indicate principals' reports of higher levels of student morale and commitment. Item wording and parameters are presented in Table 17.47.

A CFA of the two-factor model confirm the dimensionality of these items. However, the model fit is only moderate (RMSEA = 0.086, RMR = 0.013, CFI = 0.92, NNFI = 0.90) for the international school calibration sample. The estimated correlations between the two latent constructs is 0.59. Table 17.48 shows the reliabilities for the two indices, which are satisfactory in most participating countries.

Factors affecting school climate

The index of school principals' perceptions of teacher-related factors affecting school climate (TEACBEHA) is derived from seven items measuring the school principal's perceptions of potential factors hindering the learning of students at school. All items were inverted for scaling and positive scores indicate positive evaluations of this aspect. Item wording and IRT model parameters are shown in Table 17.49.

Table 17.48 ■ **Reliabilities for indices on morale and commitment**

		STMORALE	TCMORALE
OECD countries	**Australia**	0.92	0.85
	Austria	0.82	0.80
	Belgium	0.76	0.76
	Canada	0.86	0.85
	Czech Republic	0.62	0.64
	Denmark	0.81	0.72
	Finland	0.78	0.66
	France	N/A	N/A
	Germany	0.81	0.77
	Greece	0.87	0.82
	Hungary	0.81	0.77
	Iceland	0.87	0.82
	Ireland	0.81	0.84
	Italy	0.80	0.79
	Japan	0.93	0.81
	Korea	0.89	0.82
	Luxembourg	0.54	0.62
	Mexico	0.85	0.87
	Netherlands	0.75	0.75
	New Zealand	0.88	0.78
	Norway	0.76	0.81
	Poland	0.80	0.75
	Portugal	0.78	0.65
	Slovak Republic	0.80	0.74
	Spain	0.74	0.73
	Sweden	0.83	0.69
	Switzerland	0.72	0.73
	Turkey	0.91	0.90
	United Kingdom	0.91	0.83
	United States	0.85	0.88
	OECD median	**0.81**	**0.78**
Partner countries	**Brazil**	0.86	0.85
	Hong Kong-China	0.84	0.79
	Indonesia	0.90	0.86
	Latvia	0.64	0.75
	Liechtenstein	0.56	0.87
	Macao-China	0.82	0.85
	Russian Federation	0.75	0.74
	Serbia	0.85	0.70
	Thailand	0.91	0.91
	Tunisia	0.89	0.86
	Uruguay	0.80	0.77

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

Table 17.49 ■ **Item parameters for teacher-related factors affecting school climate (TEACBEHA)**

	In your school, to what extent is the learning of students hindered by:	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST25Q01	a) Teachers' low expectations of students?	-0.19	-2.64	0.40	2.25
ST25Q03	c) Poor student-teacher relations?	-0.02	-2.31	-0.49	2.81
ST25Q05	e) Teachers not meeting individual students' needs?	0.50	-2.93	0.10	2.84
ST25Q06	f) Teacher absenteeism?	0.09	-1.99	-0.43	2.42
ST25Q09	i) Staff resisting change?	0.23	-2.39	0.02	2.38
ST25Q11	k) Teachers being too strict with students?	-0.72	-2.20	-0.64	2.84
ST25Q13	m) Students not being encouraged to achieve their full potential?	0.12	-2.10	-0.06	2.16

Note: Categories were "not at all", "very little", "to some extent" and "a lot". All items were inverted for scaling.

The index on school principals' perceptions of student-related factors affecting school climate (STUDBEHA) is derived from six items measuring the school principal's perceptions of potential factors hindering the learning of students at school. All items were inverted for scaling so that positive scores indicate positive evaluations of this aspect. Table 17.50 shows item wording and IRT model parameters.

Table 17.50 ■ **Item parameters for student-related factors affecting school climate (STUDBEHA)**

	In your school, to what extent is the learning of students hindered by:	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
ST25Q02	b) Student absenteeism?	1.06	-2.52	0.07	2.45
ST25Q04	d) Disruption of classes by students?	0.69	-2.91	0.02	2.89
ST25Q07	g) Students skipping classes?	0.13	-2.29	-0.15	2.44
ST25Q08	h) Students lacking respect for teachers?	-0.18	-2.65	-0.29	2.94
ST25Q10	j) Student use of alcohol or illegal drugs?	-1.04	-1.24	-0.61	1.85
ST25Q12	l) Students intimidating or bullying other students?	-0.67	-2.46	-0.45	2.91

Note: Categories were "not at all", "very little", "to some extent" and "a lot". All items were inverted for scaling.

The index of school principals' perceptions of teacher consensus on mathematics teaching (TCHCONS) for PISA 2003 was derived from three items asking about the school principal's view of having frequent disagreement among teachers regarding (i) innovation, (ii) teacher expectations and (iii) teaching goals. The items were not inverted for scaling so that positive scores indicate higher levels of consensus among teachers (less frequent disagreements). Table 17.51 contains item wording and the IRT model parameters used for the scaling of these items.

Table 17.51 ■ **Item parameters for teacher consensus (TCCONS)**

	How much do you agree with these statements about:	Parameter estimates			
		Delta	Tau(1)	Tau(2)	Tau(3)
	Innovation in your school?				
ST21Q03	c) There are frequent disagreements between "innovative" and "traditional" mathematics teachers.	0.33	-3.56	-0.45	4.01
	Teachers' expectations in your school?				
ST22Q03	c) There are frequent disagreements between mathematics teachers who consider each other to be "too demanding" or "too lax".	0.02	-3.59	-0.56	4.15
	Teaching goals in your school?				
ST23Q03	c) There are frequent disagreements between mathematics teachers who consider each other as "too focused on skill acquisition" or "too focused on the affective development" of the student.	-0.35	-3.81	-0.82	4.64

Note: Item categories were "strongly agree", "agree", "disagree" and "strongly disagree". Items were not inverted for scaling.

The CFA results for a three-factor model show a satisfactory fit (RMSEA = 0.076, RMR = 0.028, CFI = 0.91, NNFI = 0.89) for the international school calibration sample and the estimated correlations between constructs are 0.75 for TEACBEHA and STUDBEHA, 0.11 for TCHCONS and STUDBEHA and 0.35 for TEACBEHA and TCHCONS. Table 17.52 shows that the reliabilities for these three indices were mostly satisfactory in participating countries.

School management indices

As in PISA 2000, school principals were asked to indicate who has the main responsibility for different types of decisions regarding the management of the school. The wording was only slightly modified for PISA 2003, and two comparable indices were obtained:

- Index of school autonomy (SCHAUTON): Responses indicating that decision making was not a school responsibility (first column) were recoded to 0 and those with ticks in other

Table 17.52 ■ **Reliabilities for indices of student and teacher-related factors affecting school climate and mathematics teacher consensus**

		STUDBEHA	TEACBEHA	TCHCONS
OECD countries	**Australia**	0.86	0.82	0.89
	Austria	0.77	0.77	0.83
	Belgium	0.88	0.79	0.77
	Canada	0.85	0.83	0.80
	Czech Republic	0.67	0.72	0.80
	Denmark	0.76	0.78	0.72
	Finland	0.67	0.70	0.83
	France	N/A	N/A	N/A
	Germany	0.85	0.68	0.81
	Greece	0.91	0.95	0.73
	Hungary	0.85	0.83	0.75
	Iceland	0.81	0.81	0.85
	Ireland	0.87	0.79	0.75
	Italy	0.77	0.83	0.82
	Japan	0.83	0.80	0.75
	Korea	0.90	0.85	0.84
	Luxembourg	0.77	0.76	0.74
	Mexico	0.77	0.87	0.87
	Netherlands	0.78	0.77	0.82
	New Zealand	0.85	0.82	0.80
	Norway	0.75	0.76	0.62
	Poland	0.79	0.80	0.75
	Portugal	0.78	0.75	0.76
	Slovak Republic	0.68	0.77	0.84
	Spain	0.81	0.83	0.82
	Sweden	0.81	0.79	0.76
	Switzerland	0.76	0.73	0.80
	Turkey	0.85	0.81	0.75
	United Kingdom	0.88	0.83	0.78
	United States	0.77	0.75	0.77
	OECD median	**0.81**	**0.79**	**0.80**
Partner countries	**Brazil**	0.88	0.88	0.77
	Hong Kong-China	0.94	0.92	0.76
	Indonesia	0.90	0.92	0.76
	Latvia	0.77	0.82	0.83
	Liechtenstein	0.68	0.80	0.89
	Macao-China	0.95	0.92	0.67
	Russian Federation	0.88	0.82	0.74
	Serbia	0.82	0.82	0.67
	Thailand	0.79	0.81	0.88
	Tunisia	0.89	0.81	0.51
	Uruguay	0.76	0.84	0.76

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

columns but not in the first were recoded to 1. The resulting 12 items were scaled using IRT and positive scores indicate higher levels of school autonomy in decision making.

- Index on teacher participation (TCPARTI): Responses with a tick in the last column (indicating that teacher have a main responsibility) were recoded to 1, responses with no tick but ticks in other columns to 0. The resulting 12 items were scaled using IRT and positive scores indicate higher levels of teacher participation in decision making.

Table 17.53 shows the item wording and the item parameters for both indices.

Table 17.53 ■ **Item parameters for school management indices**

	In your school, who has the main responsibility for:	Item parameters for scale	
		SCHAUTON	TCHPARTI
SC26Q01	a) Selecting teachers for hire?	0.60	1.67
SC26Q02	b) Firing teachers?	1.08	3.69
SC26Q03	c) Establishing teachers' starting salaries?	3.43	4.20
SC26Q04	d) Determining teachers' salary increases?	3.29	3.76
SC26Q05	e) Formulating the school budget?	-0.04	1.33
SC26Q06	f) Deciding on budget allocations within the school?	-2.49	0.30
SC26Q07	g) Establishing student disciplinary policies?	-2.14	-2.75
SC26Q08	h) Establishing student assessment policies?	-1.31	-3.40
SC26Q09	i) Approving students for admittance to the school?	-0.81	0.73
SC26Q10	j) Choosing which textbooks are used?	-1.92	-4.34
SC26Q11	k) Determining course content?	0.32	-3.41
SC26Q12	l) Deciding which courses are offered?	-0.02	-1.77

Note: Categories were "tick", "no tick". For SCHAUTON, "no tick" was coded as 1, "tick" as 0, for TCHPARTI "tick" was coded as 1, "no tick" as 0.

Table 17.54 shows the scale reliabilities for the two indices across participating countries. Whereas SCHAUTON has satisfactory internal consistency in most national samples, TCHPARTI has rather low reliability in a number of countries.

The index of economic, social and cultural status (ESCS)

Computation of ESCS

The ESCS index for PISA 2003 was derived from three variables related to family background: highest level of parental education (in number of years of education according to the ISCED classification), highest parental occupation (HISEI score) and number of home possessions (WLEs).

The rationale for using these three components is that socio-economic status is usually seen as based on education, occupational status and income. As no direct income measure is available from the PISA data, the existence of household items is used as an approximate measure of family wealth.

Missing values for students with one missing response and two valid responses were imputed with predicted values plus a random component based on a regression of the variable with missing responses on the other two variables. Variables with imputed values were then transformed to an international metric with OECD averages of 0 and OECD standard deviations of 1. These OECD standardised variables were used for a principal component analysis applying an OECD population weight giving each OECD country a weight of 1000.

The ESCS scores were obtained as factor scores for the first principal component with 0 being the score of an average OECD student and 1 the standard deviation across equally weighted OECD countries. For partner countries, ESCS scores were obtained as

Table 17.54 ■ **Reliabilities for indices of student and teacher-related factors affecting school climate**

		SCHAUTON	TCHPARTI
OECD countries	Australia	0.68	0.74
	Austria	0.45	0.64
	Belgium	0.76	0.59
	Canada	0.76	0.65
	Czech Republic	0.78	0.65
	Denmark	0.65	0.71
	Finland	0.67	0.64
	France	N/A	N/A
	Germany	0.63	0.64
	Greece	0.81	0.85
	Hungary	0.60	0.64
	Iceland	0.53	0.55
	Ireland	0.63	0.54
	Italy	0.72	0.51
	Japan	0.81	0.84
	Korea	0.56	0.66
	Luxembourg	0.94	N/A
	Mexico	0.90	0.66
	Netherlands	0.81	0.58
	New Zealand	0.38	0.77
	Norway	0.54	0.67
	Poland	0.52	0.48
	Portugal	0.77	0.53
	Slovak Republic	0.72	0.63
	Spain	0.73	0.75
	Sweden	0.65	0.76
	Switzerland	0.74	0.71
	Turkey	0.71	0.68
	United Kingdom	0.71	0.80
	United States	0.80	0.75
	OECD median	**0.71**	**0.66**
Partner countries	Brazil	0.79	0.58
	Hong Kong-China	0.63	0.73
	Indonesia	0.82	0.57
	Latvia	0.65	0.62
	Liechtenstein	0.85	0.79
	Macao-China	0.80	0.86
	Russian Federation	0.66	0.64
	Serbia	0.77	0.45
	Thailand	0.68	0.85
	Tunisia	0.45	0.50
	Uruguay	0.92	0.59

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.

$$ESCS = \frac{\lambda_1 HISEI' + \lambda_2 PARED' + \lambda_3 HOMEPOS'}{\varepsilon_f} \qquad (17.5)$$

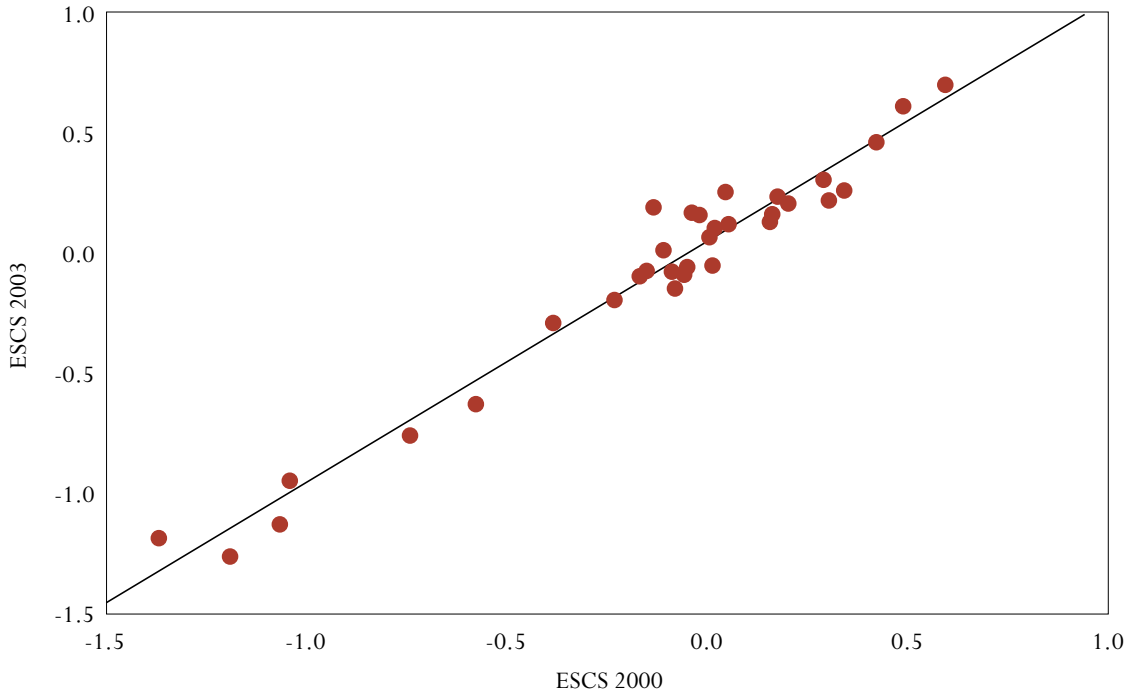
where λ_1, λ_2 and λ_3 are the OECD factor loadings, *HISEI'*, *PARED'* and *HOMEPOS'* the OECD-standardised variables and ε_f is the eigenvalue of the first principal component.[9]

Consistency between PISA 2000 and PISA 2003

ESCS was computed for PISA 2003 and also re-computed for the PISA 2000 data. There were some deviations as parental education in PISA 2000 had only one combined category for ISCED 5A and 5B whereas completion of these levels was asked separately in 2003. Furthermore, the index of home possessions was based on the subset of 11 household items that were common in both cycles. Comparing ESCS mean scores per country shows that in spite of these differences there is a very high correlation of 0.98 between ESCS 2000 and ESCS 2003 country means. Figure 17.3 shows the weighted averages for countries participating in both cycles.

In PISA 2000, a different ESCS index was used, which had been derived from occupational status of parents, parental education and the home background indices on cultural possessions (CULTPOSS), home educational resources (HEDRES) and family wealth (WEALTH) (OECD, 2001). The country-level correlation between the previous ESCS (based on five components) and the re-computed ESCS (based on three components) is around 0.94 within PISA 2000 country samples. At the country level, both indices are correlated at 0.95.

Figure 17.3 ■ **Scatter plot of country means for ESCS 2000 and ESCS 2003**



Consistency across countries

Using principal component analysis (PCA) to derive factor loading for each participating country provides insight into the extent to which there are similar relationships between the three components. Table 17.55 shows the PCA results for the 40 participating countries and the scale reliabilities for the z-standardised variables.

Comparing results from within-country PCA reveals that patterns of factor loadings are generally similar across countries. All three components contribute more or less equally to this index with factor loadings

ranging from 0.65 to 0.85. Internal consistency ranges between 0.56 and 0.77, the scale reliability for the pooled OECD sample with equally weighted country data is 0.69.

Table 17.55 ■ **Factor loadings and internal consistency of ESCS 2003**

		Factor loadings			
		HISEI	PARED	HOMEPOS	Reliability[1]
OECD countries	Australia	0.79	0.75	0.70	0.61
	Austria	0.79	0.78	0.74	0.65
	Belgium	0.82	0.77	0.74	0.67
	Canada	0.79	0.79	0.68	0.62
	Czech Republic	0.86	0.84	0.70	0.72
	Denmark	0.79	0.76	0.73	0.63
	Finland	0.80	0.75	0.69	0.60
	France	0.83	0.79	0.75	0.70
	Germany	0.80	0.76	0.74	0.65
	Greece	0.85	0.81	0.76	0.73
	Hungary	0.84	0.87	0.77	0.76
	Iceland	0.75	0.79	0.65	0.56
	Ireland	0.80	0.80	0.73	0.67
	Italy	0.84	0.82	0.72	0.71
	Japan	0.72	0.76	0.70	0.56
	Korea	0.75	0.77	0.73	0.61
	Luxembourg	0.83	0.80	0.72	0.69
	Mexico	0.80	0.82	0.78	0.72
	Netherlands	0.81	0.78	0.73	0.66
	New Zealand	0.76	0.73	0.74	0.59
	Norway	0.80	0.76	0.70	0.62
	Poland	0.85	0.81	0.77	0.74
	Portugal	0.85	0.83	0.80	0.77
	Slovak Republic	0.82	0.79	0.71	0.67
	Spain	0.82	0.79	0.75	0.69
	Sweden	0.77	0.73	0.73	0.60
	Switzerland	0.79	0.77	0.73	0.64
	Turkey	0.81	0.85	0.79	0.76
	United Kingdom	0.79	0.76	0.74	0.65
	United States	0.76	0.78	0.74	0.63
	OECD parameters	**0.81**	**0.80**	**0.76**	**0.69**
Partner countries	Brazil	0.82	0.77	0.74	0.67
	Hong Kong-China	0.81	0.81	0.66	0.64
	Indonesia	0.79	0.80	0.70	0.64
	Latvia	0.78	0.70	0.71	0.56
	Liechtenstein	0.76	0.78	0.75	0.64
	Macao-China	0.77	0.77	0.70	0.61
	Russian Federation	0.82	0.82	0.68	0.67
	Serbia	0.82	0.82	0.73	0.70
	Thailand	0.85	0.87	0.76	0.77
	Tunisia	0.84	0.84	0.80	0.77
	Uruguay	0.81	0.81	0.78	0.72

Note: Reliabilities (Cronbach's alpha) computed with weighted national samples.
1. Standardised Cronbach's alpha.

Notes

1 Students who were born abroad but had at least one parent born in the country of test were also classified as "native students".

2 Data on public/private school ownership in Australia are not included in the PISA 2003 database.

3 An alternative is the rating scale model (RSM) which has the same step parameters for all items in a scale (see Andersen, 1997).

4 A similar approach was used in the IEA Civic Education Study (see Schulz, 2004).

5 The corresponding estimates from a multidimensional IRT model with ACER ConQuest were 0.41, 0.64 and 0.76 respectively.

6 Estimates for the United States were not available as items measuring BELONG were not included in its national survey.

7 In some cases estimates of latent correlations in CFA can be greater than 1.

8 Due to the modifications in the item format and wording the index is not entirely comparable to the one used in PISA 2000 (OECD, 2002).

9 Only one principal component with an eigenvalue greater than one was identified in each of the participating countries.

18

International Database

FILES IN THE DATABASE

The PISA international database consists of three data files: two student-level files and one school-level file. All are provided in text (or ASCII) format with the corresponding SAS and SPSS control files.

The student questionnaire file

Student performance data file (filename: int_stui_2003.txt)

For each student who participated in the assessment, the following information is available:

- Identification variables for the country, adjudicated sub-national region, stratum, school and student;
- The student responses on the three questionnaires, *i.e.* the student questionnaire and the two international options: Information Communication Technology (ICT) questionnaire and Educational Career (EC) questionnaire;
- The students' indices derived from the original questions in the questionnaires;
- The students' performance scores in mathematics, reading, science and problem solving; and
- The students' weights and 80 Fay's replicates for the computation of the sampling variance estimates.

The assessment items data file

The cognitive file (filename: int_cogn_2003.txt)

For each student who participated in the assessment, the following information is available:

- Identification variables for the country, school and student; and
- The students' responses for each item included in the test expressed in a one-digit format.[1]

The school file

The school questionnaire data file (filename: int_schi_2003.txt)

For each school that participated in the assessment, the following information is available:

- Identification variables for the country, adjudicated sub-national region, stratum and school;
- The school responses on the school questionnaire;
- The school indices derived from the original questions in the school questionnaire; and
- The school weight.

RECORDS IN THE DATABASE

Records included in the database

Student level

- All PISA students who attended test (assessment) sessions.
- PISA students who only attended the questionnaire session are included if they provided a response to the father's occupation questions or the mother's occupation questions on the student questionnaire (questions 7 to 10).

School level

- All participating schools – that is, any school where at least 25 per cent of the sampled eligible students were assessed – have a record in the school-level international database, regardless of whether the school returned the school questionnaire.

Records excluded from the database

Student level

- Additional data collected by some countries for a national or international option such as a grade sample.
- Sampled students who were reported as not eligible, students who were no longer at school, students who were excluded for physical, intellectual or language reasons, and students who were absent on the testing day.
- Students who refused to participate in the assessment sessions.
- Students from schools where less than 25 per cent of the sampled and eligible students participated.

School level

- Schools where fewer than 25 per cent of the sampled eligible students participated in the testing sessions.

REPRESENTING MISSING DATA

The coding of the data distinguishes between four different types of missing data:

- *Item level non-response*: 9 for a one-digit variable, 99 for a two-digit variable, 999 for a three-digit variable, and so on. Missing codes are shown in the codebooks. This missing code is used if the student or school principal was expected to answer a question, but no response was actually provided.
- *Multiple or invalid responses*: 8 for a one-digit variable, 98 for a two-digit variable, 998 for a three-digit variable, and so on. For the multiple-choice items code 8 is used when the student selected more than one alternative answer.
- *Not-applicable*: 7 for a one-digit variable, 97 for a two-digit variables, 997 for a three-digit variable, and so on for the student questionnaire data file and for the school data file. Code 'n' instead of 7 is used for a one-digit variable in the test booklet data file. This code is used for cognitive item and questionnaire items that were not administered to the students and if a question was misprinted or deleted from the questionnaire by a national centre.
- *Not reached items*: all consecutive missing values clustered at the end of test session were replaced by the non-reached code, 'r', except for the first value of the missing series, which is coded as missing.

HOW ARE STUDENTS AND SCHOOLS IDENTIFIED?

The student identification from the student files consists of three variables, which together form a unique identifier for each student:

- The country identification variable labelled COUNTRY. The country codes used in PISA are the ISO 3166 country codes.

- The school identification variable labelled SCHOOLID. These are sequential numbers, which were randomly assigned for confidentiality reasons.
- The student identification variable labelled STIDSTD. These are sequential numbers, which were randomly assigned for confidentiality reasons.

A fourth variable has been included to differentiate adjudicated sub-national entities within countries. This variable (SUBNATIO) is used for three countries as follows:

- *Italy* The value '01' is assigned to the region 'Veneto-Nord-Est', '02' to the region Trento-Nord-Est region, '03' to the region 'Toscana-Centro', '04' to the region 'Piemonte-Nord-Ovest', '05' to the region 'Lombardia-Nord Ovest', '06' to the region 'Bolzano' and the value '07' to all other (non-adjudicated) Italian regions.
- *Spain* The value '01' is assigned to the non-adjudicated regions in Spain, '02' to Castilia and Leon, '03' to Catalonia and '04' is assigned to Basque Country.
- *United Kingdom* The value '01' is assigned to England, Northern Ireland and Wales and the value '02' is assigned to Scotland.

A fifth variable (STRATUM) contains information on the explicit strata used for sampling. Some of these were combined into larger units for policy or confidentiality reasons.

The school identification consists of two variables, which together form a unique identifier for each school:[2]

- The country identification variable labelled COUNTRY. The country codes used in PISA are the ISO 3166 country codes.
- The school identification variable labelled SCHOOLID.

THE STUDENT QUESTIONNAIRE FILE

Two types of indices are provided in the student questionnaire files. The first type is based on a transformation of one variable, or on a combination of the information included in two or more variables. Twenty indices of this first type are included in the database. The second type is scaled indices with weighted likelihood estimates as individual scores. Nineteen scaled indices from the student questionnaire and six scaled indices from the international option on Information Communication Technology are included in the database. In addition, the index of economic, social and cultural status is available in this data file. For a full detailed description of the student-level indices, including their construction and validation, see Chapter 17. In the international data files, the variable W_FSTUWT is the final student weight. The sum of the weights constitutes an estimate of the size of the target population, *i.e.* the number of 15-year-old students in grade 7 or above attending school in that country.

Note that if an analysis were performed at the international level then large countries would have a stronger contribution to the results than small countries. Two country adjustment factors are included in the file to deal with this:

- CNTFAC1 can be used for the computation of equal country weights. The weight W_FSTUWT*CNTFAC1 will give an equal weight of 1000 cases to each country so that smaller and larger countries contribute equally to the analysis. In order to obtain weights with equally weighted OECD countries, one

needs to add the variable OECD indicating country membership as an additional multiplier (W_FSTUWT*CNTFAC1*OECD).

- CNTFAC2 allows the computation of normalised or standardised weights, which should be used in multi-level analysis. The weight W_FSTUWT*CNTFAC2 will give countries weights according to their sample sizes so that the sum of weights in each country is equal to the number of students in the database.

A set of five plausible values transformed to the international PISA metric is provided for each domain, *i.e.* mathematics, reading, science and problem solving, and for each subscale in mathematics (change and relationships, space and shape, quantity and uncertainty).

Mathematics and problem-solving plausible values were transformed to the PISA scale using the data for participating PISA 2003 OECD countries. This linear transformation used weighted data, with an additional adjustment factor so that each country contributes equally in the computation of the standardisation parameters.

The weighted average of five means and five standard deviations for each scale is 500 and 100 respectively for the OECD countries, but the means and standard deviations of the individual plausible values are not exactly 500 and 100, respectively. The same transformation as for mathematics was applied to the four mathematics subscales.

Science and reading plausible values were mapped to the PISA 2000 scale and then the PISA 2000 transformation, that gives OECD mean 500 and standard deviation of 100 to the reading and science scales in PISA 2000 was applied to PISA 2003 plausible values.

For a full description of the weighting methodology, including the test design, calculation of the reading weights, adjustment factors and how to use the weights, see the *PISA 2003 Data Analysis Manual: SAS® Users* (OECD, 2005a) and the *PISA 2003 Data Analysis Manual: SPSS® Users* (OECD, 2005b).

THE COGNITIVE FILES

The file with the test data (filename: int_cogn_2003.txt) contains individual students' responses to all items used for the international item calibration and in the generation of the plausible values. All item responses included in this file have a one-digit format, which contains the coded response of the student on that item.

The PISA items are organised into units. Each unit consists of a piece of text or related texts, followed by one or more questions. Each unit is identified by a short label and by a long label. The units' short labels consist of four characters. The first character is R, M, S or X for reading, mathematics, science or problem solving respectively. The three next characters indicate the unit name. For example, R055 is a reading unit called *Drugged Spiders*. The full item label (usually seven-digit) represents each question within a unit. Thus items within a unit have the same initial four characters: all items in the unit *Drugged Spiders* begin with 'R055', plus a question number: for example, the third question in the *Drugged Spiders* unit is R055Q03.

In the cognitive files, the items are sorted by domain, and alphabetically by the short label within domain. This means that the mathematics items appear at the beginning of the file, followed by the reading items, the science items and then the problem-solving items. Within domains, units with smaller numeric

identification appear before those with larger identification numbers, and within each unit, the first question will precede the second, and so on.

THE SCHOOL FILE

The school files contain the original variables collected through the school context questionnaire.

Two types of indices are provided in the school questionnaire files. The first set contains simple indices based on a transformation of one variable, or on a combination of two or more variables. The database includes 20 simple indices. The second set is the result of scaled indices with weighted likelihood estimate as individual scores. 10 scale indices are included in the database. For a full description of the indices and how to interpret them see the PISA 2003 data analysis manuals (OECD, 2005a and OECD, 2005b). The school base weight (SCWEIGHT), which has been adjusted for school non-response, is provided at the end of the school file.

It is possible to analyse school data using the school weight. However, due to the specific age-based sampling design of PISA, it is generally recommended to analyse the school data at the student level. In order to do this, the school data need to be merged with the student data file, adding the school records to each of the students from this school. When analysing school data at the student level, results need to be interpreted at the student level. For example, when analysing school ownership, one would not estimate the percentages of private schools and public schools, but rather the percentages of students attending a private school and the percentages of students attending public schools.

FURTHER INFORMATION

A full description of the PISA 2003 database, and guidelines on how to analyse it in accordance with the complex methodologies used to collect and process the data, is provided in the PISA 2003 data analysis manuals (OECD, 2005a and OECD, 2005b), which are available on the PISA Web site (*www.pisa.oecd.org*).

Notes

1 The responses from open-ended items could give valuable information about students' ideas and thinking, which could be fed back into curriculum planning. For this reason, the coding guides for these items in mathematics and science were designed to include a two-digit coding so that the frequency of various types of correct and incorrect response could be recorded. The first digit was the actual score. The second digit was used to categorise the different kinds of response on the basis of the strategies used by the student to answer the item. A file including these codes was available to national centres. The international database includes only the first digit.

2 This file also contains variables identifying adjudicated regions and explicit sampling strata.

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*, OCED, 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.

Appendix 1

SAMPLING FORMS

PISA 2003 SAMPLING FORM 1 MAIN STUDY TIME OF TESTING AND AGE DEFINITION

See Section 3.1 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

1. Beginning and ending dates of assessment _____ 2003 to _____ 2003

   The main study testing period must be no more than six weeks long, and between 1 March 2003 and 31 August 2003.

2. Please confirm that the assessment start date is after the first three months of the academic year.

   ☐ Yes

   ☐ No

3. Students who will be assessed were born between _____ and _____

   DD MM YY DD MM YY

4. As part of of the PISA sampling process for your country, will you be selecting students other than those born between the dates in 3) above? (For example students in a particular grade, no matter what age.)

   ☐ No

   ☐ Yes (please describe the additional population): _____

   Relationship to TIMSS Population 2:

   5. Is your country, or any part of your country, a participant in TIMSS 2003 for Population 2?

      ☐ No (You may skip questions 6. and 7. Done.)

      ☐ Yes

   6. Is your institute responsible for conducting TIMSS in your country?

      ☐ Yes

      ☐ No

   7. Are you concerned about the possible overlap between the TIMSS and PISA main study school samples in 2003? You may choose more than one reply below.

      ☐ No, not concerned because mostly the student populations are in different schools.

      ☐ No, not concerned because we are a big country and very few schools will be in both studies if the two samples are drawn independently.

      ☐ No, not concerned because we think it is acceptable for some schools to be in both studies, and we want the best possible sample for each study.

      ☐ Yes, it is important that as **few** schools as possible be in both studies.

      ☐ Yes, it is important that as **many** schools as possible be in both studies (so that the total number of schools from the two studies is as small as possible).

PISA 2003 SAMPLING FORM 2 — NATIONAL DESIRED TARGET POPULATION

See Section 3.2 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

1. Total national population of 15-year-olds: [a]

   2.1. Total national population of 15-year-olds enrolled in educational institutions and who are in grades 5 and higher: [b]

   2.2. Total number of 15-year-olds enrolled in grades 4 and below: _____

3. If the national desired target population for your country differs from the total national population of enrolled 15-year-olds in grades 5 and higher, describe the population(s) to be omitted from the total national population of enrolled 15-year-olds in grades 5 and higher. These include distinct geographic regions of the country, or specific language groups:

_____

_____

   Total enrolment omitted from the total national population of enrolled 15-year-olds in grades 5 and higher *(corresponding to the omissions listed above)*: [c]

4. Total enrolment in the national desired target population: [d]

box [b] - box [c]

5. Percentage of coverage of nationally enrolled 15-year-olds, in the national desired target population:

(box [d] / box [b]) x 100 [e]

6. Describe your data source *(Provide copies of relevant tables)*: _____

_____

_____

PISA 2003 SAMPLING FORM 3 — NATIONAL DEFINED TARGET POPULATION

See Section 3.3 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

1. Total enrolment in the national desired target population: [a]

   From box [d] on Sampling Form 2

1. School-level exclusions (see Section 3.3.2):

Description of exclusions	Number of schools	Number of students
1 -		
2 -		
3 -		
4 -		
TOTAL		[b]

Percentage of students not covered due to school-level exclusions: _____ %

(box [b] / box [a]) x 100

3. Total enrolment in national defined target population before within-school exclusions: *box [a] - box [b]* [c]

4. Anticipated within-school exclusions (students who could not be included in the PISA assessment, from schools where some students could be included):

Description of exclusions	Expected number of students
Functionally disabled students	
Intellectually disabled students	
Students with limited proficiency in test language	
Other	
TOTAL	[d]

Expected percentage of students not covered due to within-school exclusions: *(box [d] / box [a]) x 100* _____ %

5. Total enrolment in national defined target population: box [a] – (box [b] + box [d]) [e]

6. Coverage of national desired target population: (box [e] / box [a]) x 100 [f]

7. Describe your data source (Provide copies of relevant tables): _____

PISA 2003 SAMPLING FORM 4 — SAMPLING FRAME DESCRIPTION

See Sections 5.2 - 5.4 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

1. Will a sampling frame of geographic areas be used?
   - ☐ Yes — Go to 2
   - ☐ No — Go to 5
2. Specify the PSU Measure of Size to be used.
   - ☐ 15-year-old student enrolment
   - ☐ Total student enrolment
   - ☐ Number of schools
   - ☐ Population size
   - ☐ Other (please describe): _____
3. Specify the school year for which enrolment data will be used for the PSU Measure of Size: _____
4. Please provide a preliminary description of the information available to construct the area frame. Please consult with Westat for support and advice in the construction and use of an area-level sampling frame.
5. Specify the school estimate of enrolment (ENR) of 15-year-olds that will be used.
   - ☐ 15-year-old student enrolment
   - ☐ Applying known proportions of 15-year-olds to corresponding grade level enrolments
   - ☐ Grade enrolment of the modal grade for 15-year-olds
   - ☐ Total student enrolment, divided by number of grades
6. Specify the year for which enrolment data will be used for school ENR. _____
7. Please describe any other type of frame, if any, that will be used. _____

PISA 2003 SAMPLING FORM 5 — EXCLUDED SCHOOLS

See Section 5.5 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

Use additional sheets if necessary

School ID	Reason for exclusion	School ENR

PISA 2003 SAMPLING FORM 6 — TREATMENT OF SMALL SCHOOLS

See Section 5.7 of School Sampling Preparation Manual.

PISA Participant:

National Project Manager:

Date this version of this form was completed:

1. Enrolment in small schools:

Type of school based on enrolment	Number of schools	Number of students	Percentage of total enrolment
Enrolment of 15-year-old students < 18			[a]
Enrolment of 15-year-old students ≥ 18 and < 35			[b]
Enrolment of 15-year-old students ≥ 35			[c]
TOTAL			**100%**

2. If the the percentage in box [a] is 1 per cent or more and the percentage in box [b] is 4 per cent or more, then an explicit stratum of moderately small schools is required, AND an explicit stratum for very small schools is required. Please see section 5.7.2 to determine an appropriate school sample allocation for these strata of moderately small and very small schools..

   box [a] >= 1% and box [b] >= 4%? ☐ Yes *or* ☐ No

   Form an explicit stratum of moderately small schools and an explicit stratum of very small schools and record this on Sampling Form 7. Done.

3. If the percentage in box [a] is 1% or more, a stratum for very small schools is needed. Please see Section 5.7.2 to determine an appropriate school sample allocation for this stratum of very small schools.

   box [a] ≥ 1%? ☐ Yes *or* ☐ No

   Form an explicit stratum of very small schools and record this on Sampling Form 7. Done.

4. If the percentage in box [a] is less than 1 per cent and the percentage in box [b] is 4 per cent or more, an explicit stratum of small schools is required, but no special stratum for very small schools is required. Please see Section 5.7.2 to determine an appropriate school sample allocation for this stratum of small schools.

   box [a] < 1% and box [b] >= 4%? ☐ Yes *or* ☐ No

   Form an explicit stratum of small schools including moderately small schools and very small schools and record this on Sampling Form 7.Go to 5.

5. **New for PISA 2003**: If the percentage of students in very small schools that have only one or two eligible students, x, is less than 0.5%, then these very small schools can be excluded from the national defined target population only if the total extent of school- level exclusions of the type mentioned in 3.3.2 remains below 0.5%. If these schools are excluded, be sure to record this exclusion on Sampling Form 3, item 2

   x < 0.5%? ☐ Yes *or* ☐ No

   Excluding very small schools with only one or two students? ☐ Yes *or* ☐ No

PISA 2003 SAMPLING FORM 7 — STRATIFICATION

See Section 5.6 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

Explicit Stratification

1. List and describe the variables used for explicit stratification.

	Explicit stratification variables	Number of levels
1		
2		
3		
4		
5		

2. Total number of explicit strata: _____

(Note: If the number of explicit strata exceeds 99, the PISA school coding scheme will not work correctly. Consult Westat and ACER.)

Implicit Stratification

3. List and describe the variables used for implicit stratification in the order in which they will be used (*i.e.* sorting of schools within explicit strata).

	Implicit stratification variables	Number of levels
1		
2		
3		
4		
5		

PISA 2003 SAMPLING FORM 8 — POPULATION COUNTS BY STRATA

See Section 5.9 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

Use additional sheets if necessary

(1)	(2)	(3)	(4)	(5)
		Population Counts		
			Students	
Explicit Strata	Implicit Strata	Schools	ENR	MOS

PISA 2003 SAMPLING FORM 9 SAMPLE ALLOCATION BY EXPLICIT STRATA

See Section 5.9 and Appendix C Section C.2 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

Use additional sheets if necessary

(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Sample Allocation		
Explicit Strata	Stratum ID	% of eligible students in population	No. of schools in population	Schools	Students	% of eligible students expected in sample

PISA 2003 SAMPLING FORM 10 **SCHOOL SAMPLE SELECTION**

See Section 5.9 and Appendix C Section C.4 of School Sampling Preparation Manual.

PISA Participant:

National Project Manager:

Date this version of this form was completed:

Explicit Stratum:

S	D	I	RN
[a] Total Measure of Size	[b] Desired Sample Size	[c] Sampling Interval	[d] Random Number

Use additional sheets if necessary

(1) Line Numbers	(2) Selection Numbers

PISA 2003 SAMPLING FORM 11 **SCHOOL SAMPLING FRAME**

See Section 5 and Appendix C Section C.5 and C.6 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

Explixit Stratum: _____

Use additional sheets if necessary

(1) School List ID	(2) Implicit Stratum	(3) ENR	(4) MOS	(5) Cumulative MOS	(6) Sample Status	(7) PISA School ID

PISA 2003 SAMPLING FORM 12 SCHOOL TRACKING FORM

See Section 6.4 of School Sampling Preparation Manual.

PISA Participant: _____

National Project Manager: _____

Date this version of this form was completed: _____

Country:													
	Original Sample				**First Replacement**				**Second Replacement**				
Participant Sample Status	**Stratum**	**PISA ID**	**School ID**	**Participation Status**	**Stratum**	**PISA ID**	**School ID**	**Participation Status**	**Stratum**	**PISA ID**	**School ID**	**Participation Status**	**Remarks**

Appendix 2

PISA CONSORTIUM AND CONSULTANTS

Australian Council of Educational Research

Ray Adams (project director of the PISA consortium)

Alla Berezner (co-ordinator data management, data analysis)

John Cresswell (science test development)

Eveline Gebhardt (data processing, data analysis)

Beatrice Halleux (translation quality control)

Marten Koomen (administration)

Dulce Lay (data processing)

Le Tu Luc (data processing)

Greg Macaskill (data processing)

Barry McCrae (mathematics and problem-solving test development)

Joy McQueen (reading test development)

Juliette Mendelovits (reading test development)

Martin Murphy (field operations and sampling)

Van Nguyen (data processing)

Alla Routitsky (data processing)

Wolfram Schulz (co-ordinator questionnaire development, data processing and data analysis)

Ross Turner (co-ordinator test development)

Maurice Walker (sampling, data processing, questionnaire development)

Margaret Wu (mathematics and problem-solving test development, data analysis)

Westat

Nancy Caldwell (director of the PISA consortium for field operations and quality monitoring)

Ming Chen (weighting)

Fran Cohen (weighting)

Susan Fuss (weighting)

Brice Hart (weighting)

Sharon Hirabayashi (weighting)

Sheila Krawchuk (sampling and weighting)

Christian Monseur (weighting)

Phu Nguyen (weighting)

Mats Nyfjall (weighting)

Merl Robinson (field operations and quality monitoring)

Keith Rust (director of the PISA consortium for sampling and weighting)

Leslie Wallace (weighting)

Erin Wilson (weighting)

Citogroep

Steven Bakker (science test development)

Truus Dekker (mathematics test development)

Janny Harmsen (office and meeting support)

Kees Lagerwaard (mathematics test development)

Gerben van Lent (mathematics test development)

Ger Limpens (mathematics test development)

Ico de Roo (science test development)

Norman Verhelst (technical advice, data analysis)

Educational Testing Service

Irwin Kirsch (reading test development)

National Institute for Educational Policy Research of Japan

Hanako Senuma (mathematics test development)

Other experts

Aletta Grisay (technical advice, data analysis, translation, questionnaire development)

Donald Hirsch (editorial review)

Peter Poole (University of Leeds, problem-solving test development)

Bronwen Swinnerton (University of Leeds, problem-solving test development)

John Threlfall (University of Leeds, problem-solving test development)

J. Douglas Willms (University of New Brunswick, questionnaire development)

Reading Expert Group

Irwin Kirsch (Chair) (Educational Testing Service, United States)

Marilyn Binkley (National Center for Educational Statistics, United States)

Alan Davies (University of Edinburgh, United Kingdom)

Stan Jones (Statistics Canada, Canada)

John de Jong (Language Testing Services, The Netherlands)

Dominique Lafontaine (Université de Liège Sart Tilman, Belgium)

Pirjo Linnakylä (University of Jyväskylä, Finland)

Martine Rémond (Institut National de Recherche Pédagogique, France)

Mathematics Expert Group

Jan de Lange (Chair) (Utrecht University, The Netherlands)

Werner Blum (Vice chair) (University of Kassel, Department of Mathematics, Germany)

Vladimir Burjan, (EXAM, Slovak Republic)

Sean Close, (St Patrick's College, Dublin, Ireland)

John Dossey (Illinois State University, United States)

Mary Lindquist (Vice chair) (Columbus College, United States)

Zbigniew Marciniak (Warsaw University, Institute of Mathematics, Poland)

Mogens Niss (Roskilde University, Denmark)

Kyung-Mee Park (Hongik University, Korea)

Luis Rico (University of Granada, Spain)

Yoshinori Shimizu (Tokyo Gakugei University, Japan)

Science Expert Group

Wynne Harlen (Chair) (University of Bristol, United Kingdom)

Peter Fensham (Monash University, Australia)

Raul Gagliardi (University of Geneva, Switzerland)

Svein Lie (University of Oslo, Norway)

Manfred Prenzel (Universität Kiel, Germany)

Senta Raizen (National Center for Improving Science Education (NCISE), United States)

Donghee Shin (Dankook University, Korea)

Elizabeth Stage (University of California, United States)

Problem Solving Expert Group

John Dossey (Chair) (Consultant, United States)

Beno Csapo (University of Szeged, Hungary)

Wynne Harlen (University of Bristol, United Kingdom)

Ton de Jong (University of Twente, The Netherlands)

Eckhard Klieme (German Institute for International Educational Research, Germany)

Irwin Kirsch (Educational Training Service, United States)

Jan De Lange (Utrecht University, The Netherlands)

Stella Vosniadou (University of Athens, Greece)

Technical Advisory Group

Geoff Masters (Chair until July 2001) (ACER, Australia)

Ray Adams (ACER, Australia)

Pierre Foy (IEA Data Processing Centre, Germany)

Aletta Grisay (Belgium)

Larry Hedges (University of Chicago, United States)

Eugene Johnson (American Institutes for Research, United States)

John de Jong (Language Testing Services, The Netherlands)

Christian Monseur (from July 2001) (HallStat SPRL, University of Liège, Belgium)

Keith Rust (Chair from July 2001) (WESTAT, USA)

Norman Verhelst (Cito group, The Netherlands)

J. Douglas Willms (University of New Brunswick, Canada)

Appendix 3

COUNTRY MEANS AND RANKS BY BOOKLET

Table A3.1

Mathematics means and ranks for each booklet containing mathematics items after application of an international booklet correction

	Booklet 1		Booklet 2		Booklet 3		Booklet 4		Booklet 5		Booklet 6		Booklet 7		Booklet 8		Booklet 9		Booklet 10		Booklet 11		Booklet 12		Booklet 13	
	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank
FIN	548	1	540	3	536	7	542	5	541	3	546	2	535	7	551	6	555	2	546	4	547	3	546	1	541	3
HKG	541	2	534	5	558	1	551	1	544	2	547	1	555	1	585	1	553	3	559	2	550	1	528	8	549	1
NLD	527	10	540	1	531	8	539	6	538	5	541	3	537	6	556	5	559	1	560	1	548	2	537	3	543	2
LIE	515	14	530	7	539	3	545	2	554	1	530	8	545	3	563	3	535	7	528	10	530	7	528	9	528	8
KOR	532	5	536	4	539	4	543	3	537	7	538	4	537	4	575	2	553	4	557	3	537	5	536	4	528	7
BEL	534	4	540	2	538	6	539	7	540	4	525	11	537	5	533	11	549	5	534	6	542	4	525	14	535	4
NZL	519	13	515	15	517	14	512	16	525	9	523	12	525	10	530	13	522	13	532	9	526	11	541	2	517	14
MAC	529	7	530	8	551	2	542	4	515	13	526	10	546	2	534	7	504	20	512	18	530	9	518	15	517	13
JPN	534	3	531	6	538	5	533	8	537	6	536	6	531	8	562	4	529	9	526	11	526	12	530	7	533	5
CAN	529	9	528	9	523	10	526	11	533	8	537	5	530	9	533	8	543	6	545	5	533	6	531	5	530	6
AUS	529	8	517	14	518	11	519	15	522	11	531	7	516	14	531	12	526	12	532	7	527	10	530	6	517	12
CHE	530	6	521	11	525	9	530	9	524	10	528	9	523	11	533	9	532	8	517	15	530	8	526	12	526	9
IRL	493	25	495	25	501	20	490	27	505	21	512	18	497	23	516	16	502	22	532	8	510	18	493	22	489	28
CZE	523	11	526	10	517	13	521	13	515	14	513	17	520	13	533	10	528	10	521	14	519	14	525	13	525	10
DNK	522	12	514	17	514	15	519	14	513	16	508	20	504	21	505	19	514	17	511	19	524	13	528	10	512	16
ISL	514	15	519	12	518	12	528	10	513	15	521	13	514	16	522	15	517	16	500	23	510	19	511	17	507	19
DEU	500	20	506	18	499	23	501	21	510	18	514	16	515	15	512	17	526	11	511	20	518	16	501	18	512	15
SWE	503	17	515	16	500	21	505	17	506	19	509	19	508	17	492	24	511	19	504	21	518	17	528	11	519	11
AUT	502	19	500	20	500	22	504	18	506	20	501	21	507	18	527	14	521	15	525	12	508	20	491	24	500	22
FRA	509	16	519	13	508	18	521	12	521	12	514	15	522	12	495	21	521	14	514	16	501	21	487	26	508	17
GBR	503	18	500	22	493	27	504	20	511	17	517	14	505	19	511	18	513	18	525	13	518	15	515	16	494	24
NOR	491	26	500	21	511	16	493	25	499	24	497	22	497	22	490	25	482	29	492	25	488	28	498	19	499	23
SVK	497	22	504	19	509	17	504	19	501	23	491	25	505	20	504	20	497	23	495	24	490	27	496	21	504	20
USA	485	28	445	35	467	32	466	31	494	26	490	26	480	29	486	26	494	24	513	17	498	22	467	30	490	27
POL	495	24	498	24	502	19	498	23	493	27	483	27	490	27	460	30	485	27	475	28	495	23	490	25	508	18
HUN	498	21	499	23	498	25	499	22	502	22	496	24	495	24	494	23	503	21	487	26	493	25	496	20	493	25
LUX	497	23	489	28	499	24	498	24	494	25	496	23	493	25	494	22	490	26	502	22	494	24	481	27	485	30
ESP	481	29	491	27	482	28	482	29	481	29	483	28	492	26	474	27	491	25	478	27	490	26	491	23	491	26
LVA	488	27	495	26	498	26	493	26	482	28	477	29	489	28	472	28	482	28	467	30	471	29	471	29	501	21
ITA	474	31	475	30	475	30	469	30	471	31	470	31	471	32	442	31	454	31	452	31	460	31	476	28	468	32
RUS	474	30	476	29	480	29	487	28	478	30	474	30	472	31	441	32	447	32	446	32	464	30	458	32	488	29
PRT	470	32	473	31	469	31	465	32	453	32	467	32	476	30	465	29	457	30	472	29	459	32	464	31	469	31
GRC	458	33	463	32	465	33	458	33	450	33	448	33	452	33	423	33	423	34	415	35	420	35	454	33	453	33
YUG	443	35	445	34	447	35	430	35	448	34	433	34	439	35	418	34	432	33	426	33	430	33	444	34	442	35
URY	448	34	457	33	449	34	444	34	427	35	425	35	430	37	368	38	373	37	373	37	410	37	438	35	442	34
THA	414	37	425	37	432	37	423	37	419	37	413	37	442	34	400	36	402	35	410	36	416	36	399	37	425	36
TUR	432	36	434	36	434	36	428	36	419	36	423	36	433	36	416	35	398	36	417	34	427	34	427	36	416	37
MEX	400	38	399	38	407	38	398	38	388	38	381	38	378	38	372	37	353	39	368	38	381	38	387	38	398	38
IDN	362	41	367	41	378	41	380	40	362	41	340	40	372	40	352	39	356	38	346	39	359	40	332	41	378	40
BRA	371	39	373	40	381	40	377	41	369	39	362	39	350	41	331	41	308	41	325	40	339	41	371	39	368	41
TUN	369	40	378	39	382	39	384	39	363	40	336	41	375	39	344	40	342	40	318	41	359	39	334	40	380	39

Table A3.2

Reading means and ranks for each booklet containing reading items after application of an international booklet correction

	Booklet 1		Booklet 2		Booklet 7		Booklet 8		Booklet 9		Booklet 10		Booklet 11	
	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank
FIN	552	1	541	3	548	1	538	3	548	1	533	2	526	2
KOR	538	2	556	1	523	5	547	2	526	3	546	1	543	1
LIE	536	3	537	6	534	2	557	1	509	10	519	5	504	16
CAN	533	5	543	2	522	6	530	6	525	4	527	3	525	3
AUS	530	7	523	10	524	4	530	5	532	2	526	4	518	6
NZL	534	4	534	8	527	3	520	9	515	6	515	8	519	5
NLD	517	11	536	7	518	8	532	4	510	8	512	9	520	4
IRL	530	6	540	4	522	7	526	7	510	9	511	10	510	12
HKG	524	9	538	5	510	12	521	8	502	11	506	12	515	8
SWE	503	18	521	11	512	11	515	12	521	5	519	6	514	9
BEL	526	8	512	15	513	9	514	13	514	7	517	7	511	10
JPN	503	19	523	9	497	21	516	10	487	26	507	11	516	7
GBR	518	10	519	12	513	10	501	19	500	12	502	14	504	15
CHE	511	14	506	18	504	17	516	11	493	18	495	21	505	14
DNK	513	13	506	17	488	24	501	18	488	24	504	13	511	11
DEU	515	12	513	13	508	14	502	15	492	21	495	20	485	25
MAC	496	22	502	20	503	19	502	17	491	22	496	17	510	13
AUT	505	17	513	14	510	13	510	14	492	20	496	18	482	27
NOR	483	27	507	16	497	22	481	26	499	13	501	15	499	17
POL	506	16	479	26	503	20	490	23	499	14	491	24	490	21
CZE	498	20	505	19	504	18	496	21	494	15	496	19	490	20
FRA	497	21	501	21	507	15	502	16	490	23	498	16	496	18
USA	508	15	497	22	505	16	482	25	493	16	492	23	492	19
LVA	476	28	485	24	488	25	495	22	493	17	483	26	483	26
HUN	484	26	481	25	487	26	490	24	492	19	482	27	486	24
ISL	494	23	494	23	493	23	500	20	488	25	494	22	481	28
ESP	464	31	472	29	477	29	479	28	483	28	486	25	486	22
GRC	436	33	430	33	454	32	454	32	481	31	479	29	486	23
PRT	487	24	454	31	487	27	468	30	483	29	474	31	458	33
LUX	484	25	477	27	482	28	474	29	483	27	475	30	477	30
SVK	468	29	476	28	471	30	481	27	466	32	471	32	468	31
ITA	467	30	460	30	464	31	465	31	481	30	482	28	479	29
TUR	440	32	436	32	436	34	434	34	437	35	438	35	452	34
URY	377	38	377	37	411	37	402	36	451	33	444	34	460	32
RUS	411	35	420	34	442	33	445	33	448	34	457	33	451	35
YUG	416	34	408	36	418	36	398	37	412	38	416	38	414	37
THA	410	36	410	35	418	35	412	35	424	36	428	36	431	36
MEX	385	37	373	38	393	39	388	38	407	39	409	39	404	38
BRA	371	40	327	41	405	38	348	41	424	37	417	37	395	40
IDN	375	39	364	39	386	40	369	39	387	41	384	40	383	41
TUN	347	41	336	40	364	41	354	40	388	40	383	41	398	39

Table A3.3

Science means and ranks for each booklet containing science items after application of an international booklet correction

	Booklet 5		Booklet 6		Booklet 7		Booklet 8		Booklet 9		Booklet 12		Booklet 13	
	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank
FIN	547	1	555	2	551	1	546	2	545	2	552	4	559	2
HKG	541	3	549	3	530	6	531	5	535	3	558	1	567	1
LIE	528	10	521	11	541	2	552	1	513	12	531	9	535	8
JPN	526	12	557	1	511	15	536	4	560	1	514	17	552	3
KOR	539	4	540	4	536	4	540	3	531	4	557	2	543	5
NLD	542	2	521	12	529	8	525	7	520	9	553	3	539	7
FRA	539	5	498	21	536	3	520	11	502	19	523	12	494	22
CZE	523	13	526	7	525	11	522	9	528	5	529	11	544	4
AUS	530	9	523	10	535	5	526	6	524	6	543	5	534	10
MAC	508	18	537	5	525	10	522	10	521	7	503	20	502	19
CAN	530	8	529	6	528	9	525	8	511	13	534	8	540	6
GBR	531	7	518	13	524	12	511	15	514	11	542	6	524	13
BEL	532	6	507	17	523	13	517	12	510	14	521	13	519	14
NZL	526	11	513	15	529	7	511	16	515	10	540	7	535	9
CHE	517	14	524	8	510	16	517	13	504	17	520	14	525	12
HUN	500	19	524	9	491	24	506	19	521	8	488	23	506	17
IRL	509	17	508	16	505	17	513	14	501	20	529	10	510	16
DEU	513	16	517	14	501	18	509	17	506	16	514	16	533	11
SWE	515	15	501	18	516	14	508	18	509	15	520	15	501	20
AUT	492	24	500	19	485	27	490	26	495	25	502	21	512	15
SVK	492	23	498	22	490	25	502	20	499	22	486	24	504	18
POL	494	22	499	20	479	31	494	24	502	18	496	22	490	24
DNK	498	20	464	33	494	19	484	30	468	32	511	18	486	26
USA	496	21	491	26	491	23	492	25	490	28	511	19	489	25
ISL	491	25	497	24	493	20	500	21	494	27	476	28	497	21
RUS	483	28	498	23	492	21	498	22	498	23	455	32	471	31
ITA	466	32	490	27	483	28	488	28	499	21	464	30	467	32
NOR	488	26	482	30	492	22	474	33	484	31	483	25	476	29
LVA	481	29	491	25	485	26	494	23	488	29	481	26	490	23
ESP	475	30	486	28	482	29	485	29	496	24	461	31	477	28
GRC	452	33	471	31	468	33	489	27	494	26	433	34	448	33
LUX	483	27	483	29	476	32	479	31	486	30	481	27	485	27
PRT	471	31	466	32	479	30	475	32	465	33	473	29	474	30
TUR	441	34	432	35	438	36	431	36	428	37	434	33	425	35
YUG	437	35	441	34	434	37	432	35	440	35	425	36	432	34
URY	425	37	416	36	449	34	453	34	453	34	381	38	373	39
THA	430	36	411	37	438	35	423	37	434	36	430	35	408	36
IDN	409	38	383	40	420	38	399	39	392	40	403	37	377	38
MEX	381	41	408	38	390	41	405	38	418	38	364	39	380	37
BRA	391	39	384	39	406	39	399	41	398	39	362	40	366	40
TUN	385	40	359	41	406	40	399	40	391	41	359	41	362	41

Table A3.4

Problem-solving means and ranks for each booklet containing problem-solving items after application of an international booklet correction

	Booklet 3		Booklet 4		Booklet 9		Booklet 10		Booklet 11		Booklet 12		Booklet 13	
	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank
FIN	551	6	541	6	549	2	554	1	541	3	539	3	542	3
KOR	566	2	555	3	554	1	541	2	539	5	541	2	533	6
HKG	575	1	570	1	548	3	532	9	548	1	530	6	532	9
NZL	529	8	550	5	526	9	536	7	520	10	524	9	549	1
JPN	564	3	551	4	547	4	522	13	547	2	544	1	543	2
LIE	552	5	556	2	512	17	541	3	518	13	520	13	533	7
AUS	529	9	541	7	533	5	538	4	527	7	527	7	534	5
CAN	527	10	541	8	527	8	536	6	506	18	532	5	541	4
MAC	560	4	519	18	528	7	515	19	540	4	534	4	514	14
BEL	533	7	533	14	532	6	537	5	534	6	524	8	522	12
NLD	518	15	538	9	516	13	534	8	523	8	524	10	533	8
CZE	522	13	537	12	521	10	527	11	521	9	520	14	523	11
CHE	526	11	538	11	515	15	531	10	519	11	523	11	528	10
DEU	520	14	538	10	515	16	520	15	515	15	514	17	511	17
DNK	522	12	524	16	521	11	521	14	518	12	522	12	511	16
FRA	518	17	523	17	517	12	519	16	510	17	516	16	520	13
GBR	506	19	534	13	504	21	523	12	506	19	508	19	506	18
SWE	505	22	503	20	515	14	516	17	516	14	520	15	513	15
AUT	518	16	529	15	509	19	515	18	502	20	497	21	495	23
HUN	506	20	491	22	512	18	504	20	511	16	513	18	494	24
IRL	508	18	491	23	502	22	498	23	497	21	491	26	483	30
ISL	499	25	501	21	508	20	501	21	494	23	507	20	503	19
LUX	502	24	507	19	491	24	498	22	493	24	491	25	501	21
RUS	455	32	450	32	483	28	478	29	495	22	491	24	502	20
NOR	505	21	480	27	488	26	480	28	490	25	494	22	496	22
SVK	503	23	488	24	493	23	490	25	489	26	492	23	486	27
USA	475	28	487	26	474	30	494	24	471	31	477	31	482	31
POL	487	27	487	25	491	25	480	27	483	29	489	27	486	26
LVA	487	26	467	29	483	29	462	32	486	27	489	28	490	25
ESP	470	29	476	28	483	27	485	26	482	30	486	29	484	29
ITA	467	30	450	31	472	31	468	31	483	28	481	30	478	32
PRT	463	31	453	30	471	32	470	30	468	32	470	33	484	28
GRC	438	33	419	33	442	33	437	33	461	33	473	32	467	33
YUG	428	35	411	35	420	35	412	34	423	36	432	36	422	35
THA	432	34	401	36	431	34	392	37	444	34	436	35	399	37
TUR	426	36	412	34	399	37	407	35	418	37	410	37	410	36
URY	394	37	353	38	407	36	399	36	441	35	448	34	451	34
MEX	378	38	354	37	386	38	364	38	411	38	397	38	390	38
BRA	371	39	333	40	377	39	353	39	392	39	381	39	374	39
IDN	370	40	336	39	367	40	343	40	374	40	366	41	343	41
TUN	337	41	331	41	341	41	335	41	358	41	371	40	369	40

Appendix 4

ITEM SUBMISSION GUIDELINES FOR MATHEMATICS – PISA 2003

Mathematics Item Development for PISA 2003 and Item Submission Guidelines

Submission of mathematics items for PISA

All mathematics items for PISA 2003 should be prepared in accordance with the detailed advice presented in this document. The advice is provided in three discrete sections of the document:

Section 1: Overview of the item development task

This section includes a description of the scope of the task, a summary of the development process to be followed, the arrangements for submission and review of items, and the item development timeline.

Section 2: Specification of mathematics item requirements for PISA

This section includes reference to the mathematics framework, detailed guidelines on the required form of mathematics items, and a detailed discussion of factors contributing to item difficulty.

Section 3: Sample items

A small number of sample items are provided, showing the desired form and layout of the three major item types, and a sample marking guide for open-constructed response questions.

Wherever possible, before items are submitted some degree of refinement of items should have taken place at the country level based on pre-pilot activities that involve students.

Items should preferably be submitted in English, French, German, Dutch, Italian, Spanish, Russian or Japanese. Other languages may be used following negotiation with and the agreement of the consortium.

Items should be submitted as early as possible. If possible, items should be submitted progressively, as they are developed, rather than waiting until close to the submission deadline. Items received after June 2001 cannot be included in the field trial pool.

In preparing items for submission, item developers should provide the following information about each item.

- Information about the source of the item (original, or from a book or other source)
- Information about any copyright considerations (who holds the copyright, who should be contacted to seek permission to use the item, etc). This is particularly important for diagrams and graphical material.
- The classification of each item according to framework categories (including the problem domain and item context)

This information should be provided by completing the OECD/PISA mathematics item submission form as a cover sheet for each item submitted. The item submission form is provided at the end of this document.

Items should be submitted to ACER, preferably in electronic format as a Microsoft®Word® for PC document.

Addresses for item submission are:

By e-mail to *TurnerR@acer.edu.au*
Or by surface mail to
PISA Mathematics
(Attention: Ross Turner)
Australian Council for Educational Research
Private Bag 55
CAMBERWELL VICTORIA 3124
AUSTRALIA

Section 1: Overview of the item development task

Scope of work

For PISA 2003, about 130 mathematics items will be needed for the main study, some of which can be selected from the PISA 2000 item pool. The target will therefore be about double this (260 items) for the field trial in 2002. To achieve this about 390 items will be needed for the pilot stage, from which the field trial items will be selected.

The mathematics item development work will be shared between ACER in Australia, the Citogroup in the Netherlands and NIER in Japan. Items will also be submitted by mathematics expert group members, and by national experts and others in individual countries through the National Project Managers (NPMs).

The development process

In order to develop items of the best possible quality, some improvements in the item development process will be implemented for PISA 2003.

Various cognitive laboratory processes will be used in a more systematic and structured way than was the case for PISA 2000. Cognitive laboratory procedures are specialist activities that require training and expertise, and will in general be used by the consortium's test developers in their own countries. This includes the ACER, the Citogroup and the NIER teams developing mathematics items in Australia, the Netherlands and Japan respectively.

For example, a set of guidelines for cognitive walk-through (sometimes known as item panelling, or item shredding) will be implemented. All test developers will use this step at the early stages of item development, between the commencement of item writing and preparation of items for the field trial in 2002. This step will also be used later in 2002 following the field trial in the refinement of items for the main study.

At least some items will be subject to further cognitive laboratory processes (including, for example, think-aloud procedures, cognitive interviews, cognitive group interviews and cognitive comparison studies). Item development teams will use these procedure with students from their own countries in the development of selected items, in the period leading up to selection of items for field trial. In some cases these laboratory procedures will also be used later in 2002, following the field trial, in the refinement of items for the main study.

An important principle that will be followed is that more than one consortium partner will be involved the development of each mathematics item. Draft items will routinely be circulated among the relevant consortium partners, for implementation of the cognitive laboratory procedures appropriate for each item, in plenty of time to allow consideration of items being developed and comments to be taken into account. In addition, the consultation process used in the development and selection of items will draw on national expertise through the NPMs. Moreover, an even broader range of expertise will be tapped through the input of the mathematics forum.

The item development process is summarised in Figure A4.1.

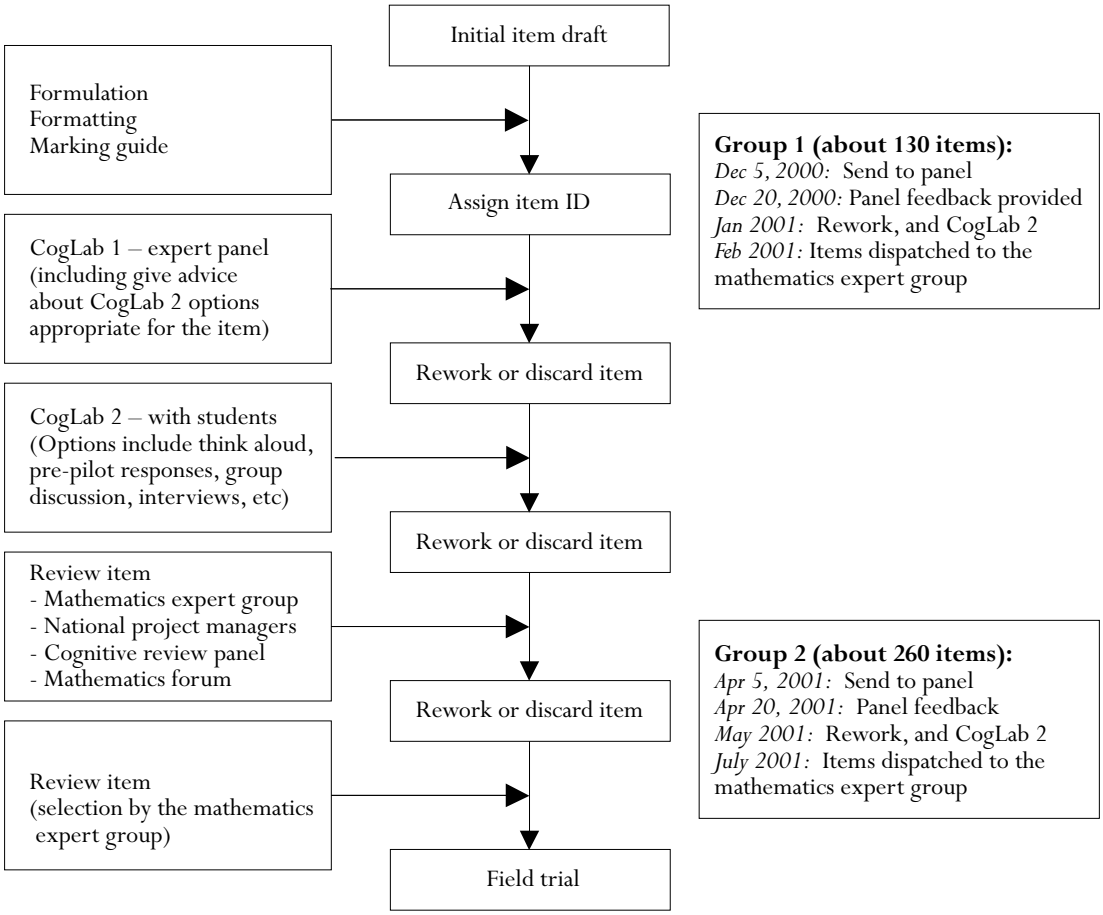
Arrangements for submission and review of items

The consortium seeks to maximise item input from mathematics experts in participating countries. This will help in ensuring a mix of items that best reflects the diversity of cultural contexts and values seen in PISA countries. These item submission guidelines have been prepared to facilitate this national input.

Wherever possible, submitted items would have gone through some refinement at the country level prior to submission. This would include complete item formulation, preparation of a draft marking guide, and a summary of the characteristics of the item from the point of view of the major classifications arising from the mathematics framework. Less well-formed input will also be welcomed and the consortium item-writing teams will carry out whatever further testing and development is required for all material submitted.

It is expected that when preparing national submissions of items, NPMs will endeavour to provide a range of items from the different problem domains described in the mathematics framework, rather than many items of a similar type from just one or two problem domains. Similarly, submitted items should vary in their format. It is also better for submissions to be smaller in size but of higher quality, than to be more extensive but of relatively poor quality.

Figure A4.1
Summary of item development process for mathematics for PISA 2003



There will also be an extensive process of consultation and review for all items developed. This process will involve the mathematics expert group, all participating countries through the NPMs, and the wider reference group, the mathematics forum. It is proposed that as bundles of items are developed, they will be circulated for comment and feedback during 2001 prior to item selection in about September 2001 for the field trial in 2002.

One additional step in the item review process will be the use of a formalised cultural review process for all items. The PISA consortium will establish an international panel of suitable experts and will routinely submit draft items to that panel for scrutiny. For field trial items, this consultation stage will be carried out in mid-2001.

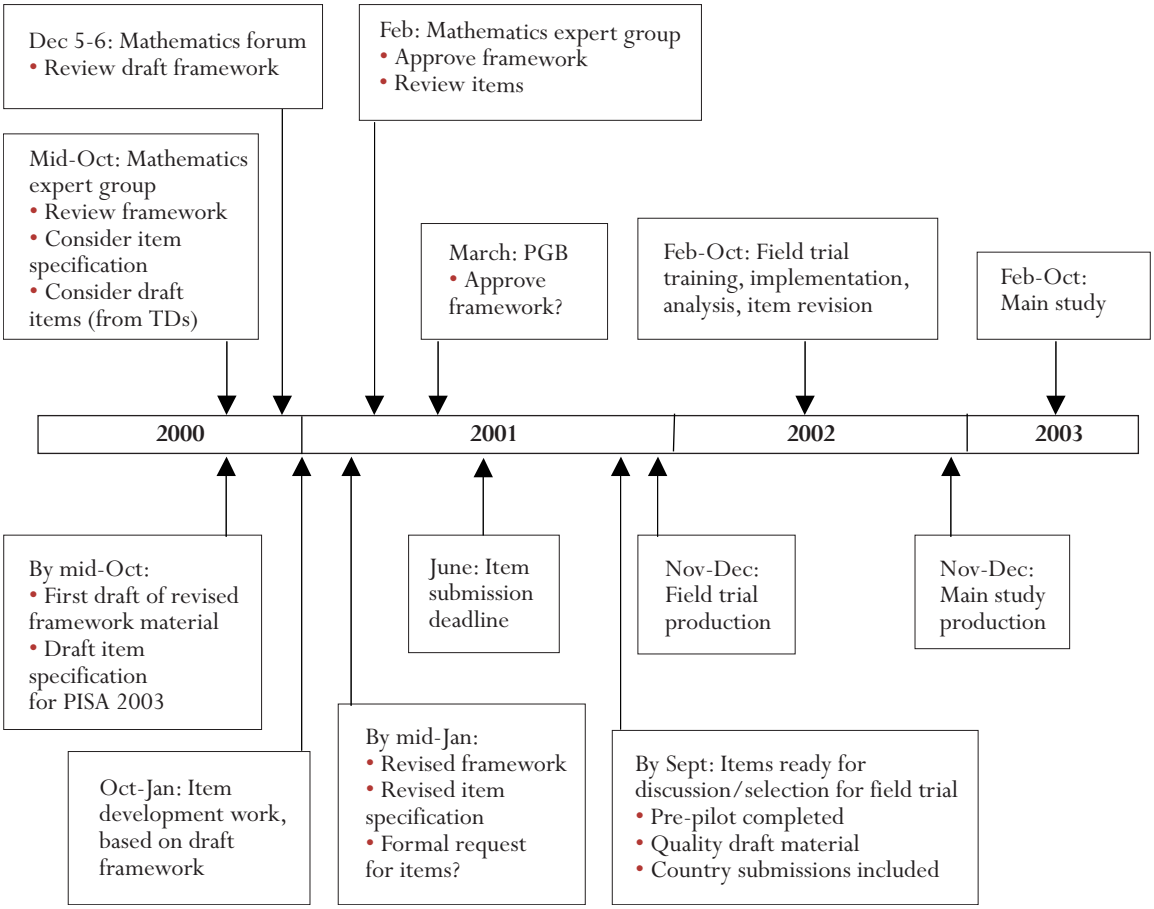
A corresponding set of review processes will also be implemented during 2002 in the lead-up to the selection of materials for the main study in 2003.

Timeline for item development

A timeline for the development of mathematics items is given in Figure A4.2.

Country submissions were invited informally, based on a draft version of the revised mathematics framework for PISA 2003, early in 2001. This more formal request is accompanied by a more complete draft of the framework.

Figure A4.2
Timeline of mathematics item development for PISA 2003



The final date for item submission is June 2001. Countries are encouraged to submit materials progressively over the first half of 2001. For material received in May or June there will be very little time to carry out any development required. Material received after June will not be used.

All development of items to the stage of selection for field trial must be complete by September 2001. Translation into French will occur in the period from September to November 2001, with distribution to countries for translation in November/December 2001 in preparation for the field trial in 2002.

Section 2: Specification of mathematics item requirements

This specification of the requirements for PISA mathematics items contains a number of elements:

- A reference to the framework within which items are to fit,
- Guidelines that describe the desired form of items to be submitted, and advice on a number of problems and pitfalls in item development that should be avoided; and
- A discussion of factors that contribute to item difficulty and the described proficiency scale for PISA mathematics.

The mathematics framework

The mathematics framework describes the way in which the mathematics domain is conceived for PISA assessment purposes. It provides the context into which all assessment items must fit. The PISA framework defines mathematical literacy as the

> Student's capacity to understand and to engage in mathematics and make well-founded judgments about the role mathematics plays, as needed for an individual's current and future private life, occupational life, social life with peers and relatives, and life as a constructive, concerned and reflective citizen.

The framework includes:

- A discussion of the broad context within which the PISA mathematics assessments occur;
- A description of the contexts and settings that should be used for items;
- The identification and definition of mathematical content areas;
- The mathematical processes underpinning the domain; and
- A description of the mathematical proficiency dimension that is assessed through PISA.

The mathematics framework will be published as a separate document during 2001.

Guidelines on the form of mathematics items

General issues

PISA is an international test of the literacy skills of 15-year-olds. All items submitted should be suitable for the population of 15-year-old students in OECD countries.

To avoid unfairly advantaging students from any particular country, items should not be directly extracted from textbooks or other common resource materials routinely used by students in any country.

In some cases, PISA items will be developed in units that comprise a number of questions (perhaps three to five) relating to a single stimulus. Such questions should as far as possible be independent of each other in the sense that the answer to one question should not depend on the student having answered a previous question.

In other cases, stand-alone items will be developed. In general, items should consist of some stimulus material or information, an introduction and the actual question, then for non-multiple-choice items a proposed marking scheme for the question.

Development of a marking scheme will be greatly assisted by using some preliminary field trial process involving students. At a minimum this might involve preparing a version of the question that is formatted for actual student use, and pilot testing the draft item with a small group of students. Information from the student responses can then be used to refine the item, and to prepare a marking guide. The consortium item development teams will carry out all required additional item refinement work for all material submitted.

Items should relate predominantly to one of the mathematical problem domains described in the framework. The link must be made explicit.

Items should be identified predominantly with one of the competency classes that are described in the framework. The links must be clear.

Items will be set in one of the situations described in the framework, and will be set in a suitable context. The item situation must be made explicit.

Approximately one-third of mathematics items should be presented in multiple-choice format. About a third will be in a form requiring a closed constructed response (that is, the response can be keyed in by a data entry operator without expert judgment, and scored automatically). The remaining third (approximately) will be in a form requiring an open constructed response (that is, requiring a detailed marking guide and the expert judgment of a trained marker). It is expected that some of the constructed response items will use a partial-credit scoring rubric. For these items, separate and detailed consideration should be given to the knowledge and skill demands of each separate score point available for the item.

The final proportions of each item type will be determined on the basis of policy decisions of the PGB, and considerations to be made at the time of item selection.

Items should require only the normal equipment that students could be expected to have, such as rulers, erasers, compasses and protractors, and calculators. While there will not be items designed specially for testing calculator skills, the PISA test will contain items for which the use of calculator may be helpful for the students. The PISA assessment focuses on problem situations that arise from the real world, and the use of calculators is very much a part of everyday life (whether at work or at home). However, it should be stressed that intensive computation is not a key focus of the PISA test, and there will not be purely computational items that depend solely on the use of the calculator.

The level of reading required to successfully engage with an item should be considered very carefully. The wording of items should be as simple and direct as possible. This is an assessment of mathematical literacy, not of reading ability.

Care should be taken to avoid question contexts that would create a cultural bias. Keep in mind that the test will be administered in a large number of countries with big cultural and geographical differences.

Testing time and test design

Items will be arranged into 30-minute blocks, with each student taking four blocks (two hours) of testing, including a mix of reading, mathematics, science and problem-solving blocks. It is recommended that each item should not require more than five minutes (for the average student) for completion. Each unit should not require more than 15 minutes for completion. Very time-consuming items should be avoided.

Item presentation and formulation

Items should be presented so that they are as clear and straight forward as possible given the required task.

Items should be expressed as far as possible in an active way rather than in a passive way. For example, "Take a number, divide it by 2 etc…." rather than "If a number is divided by 2….".

Unknown and unfamiliar words should be avoided. For example, one of the names in an item in a particular Australian test was "Nina" and quite a few children didn't do the item because they didn't know what a "Nina" was. When the name was changed to "Sue", the problem disappeared. In general this is an issue that must be handled carefully at the national level when material is translated and national versions are prepared. As PISA test material is prepared for translation, information will be provided as to which aspects of the question can and should be adapted to suit local conditions in each country, and which aspects must not be changed.

Anything that makes an item hard to comprehend should be avoided. For example, in another Australian test, an item with a female plumber in it caused many students to confuse the customer and the plumber and for this reason, students found the item very hard. Again, countries must treat such issues carefully when preparing national versions, since gender roles can differ from country to country, and the way students react to contextual factors will also differ across countries.

The order of presentation of information is another important matter. In general, PISA items should follow these principles: first the stimulus, then an introduction with the information for the question as close as possible to the question itself. If the item has several questions, information specifically for each question should be given just before that question. Lengthy and complex wording should be avoided.

Item format and the demand for marking

While item format can be multiple-choice, short-answer and extended-response, questions that would require very complex marking guides should be avoided. Countries have limited resources in employing markers. In addition, the consistency and reliability of markers across all countries need to be ensured. Therefore, it is important to have simple and objective marking guides.

The following example is too open-ended for objective scoring:

> You are going to design an aquarium made of glass, which holds about 80 litres. Suggest some appropriate measurements. Describe how you found those measurements and draw a sketch of the aquarium with your measurements.

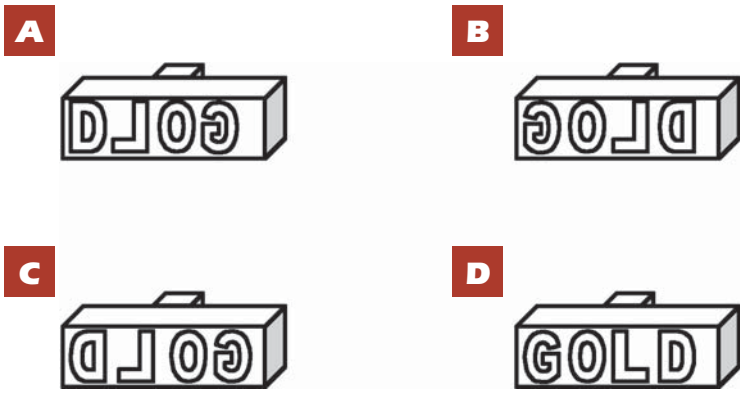
As a classroom activity, this is a good item for testing real-life problem-solving skills. But the marking scheme will be too complex, and it would be difficult to ensure consistency between markers across countries.

International equivalence – translation issues

Tests will be administered in about 40 countries. They will need to be translated into about 20 languages. Therefore, items that have specific language elements will not be suitable. An example is given below:

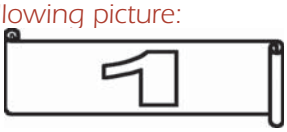
You are making a stamp to print the word 'GOLD' as follows:

Which of the following stamps will correctly print the word 'GOLD'?

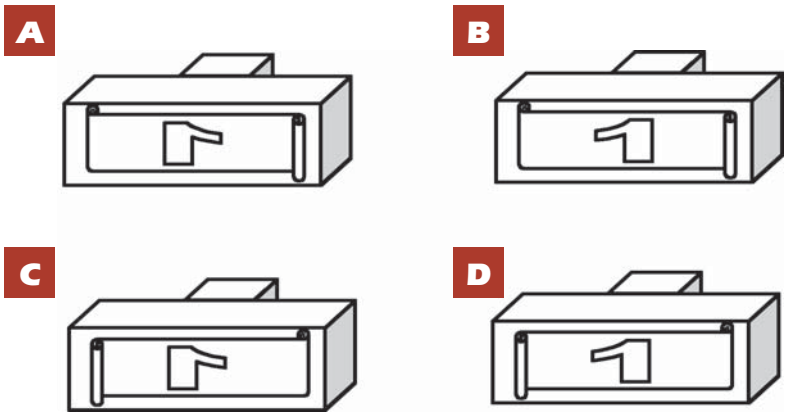


Because this item uses the letters from the Latin alphabet, it is unsuitable for students from countries such as Japan and Korea. A similar item with pictures would be better for an international test. For example, consider the following item:

You are making a stamp to print the following picture:



Which of the following stamps will correctly print that picture?



Avoid incomplete stems, as in the following example:

Line X is shorter than
A Line Y
B Line Z
C Line W
D Line T

Instead, write "Which line is longer than Line X?".

© OECD 2005

Take particular care in distinguishing between "singular" and "plural" grammatical forms, as in the following example: "Which of the following…" In English, this could mean one or more. Some languages do not allow for such ambiguity. Always explicitly state singular or plural.

Other examples of language-specific items include:

How many sides does a quadrilateral have?

In some languages, the word for quadrilateral is "four-sided figure"

Which one of the following is 45 million?

Some languages don't use the word "million".

Which letters have a line of symmetry: H,U,N,G,A,R,Y

Remember Asian languages do not use the Latin alphabet

Real-world context

The PISA mathematics framework places an emphasis on real-world contexts and authenticity of items. The following item would be regarded as too contrived:

Farmer Dave keeps chickens and rabbits. Dave counted altogether 65 heads and 180 feet. How many chickens does Dave have?

However, this question could be written as follows:

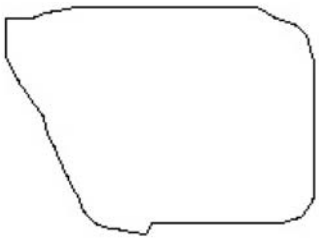
Tickets to the school concert costs 4 zeds for an adult and 2 zeds for a child. 65 tickets were sold for a total of 180 zeds. How many children's tickets were sold?

Wherever possible, use real objects rather than fictitious objects. Consider the following two items *Field Area* and *Continent Area*. *Continent Area* is a better item because of the use of a real map.

FIELD AREA

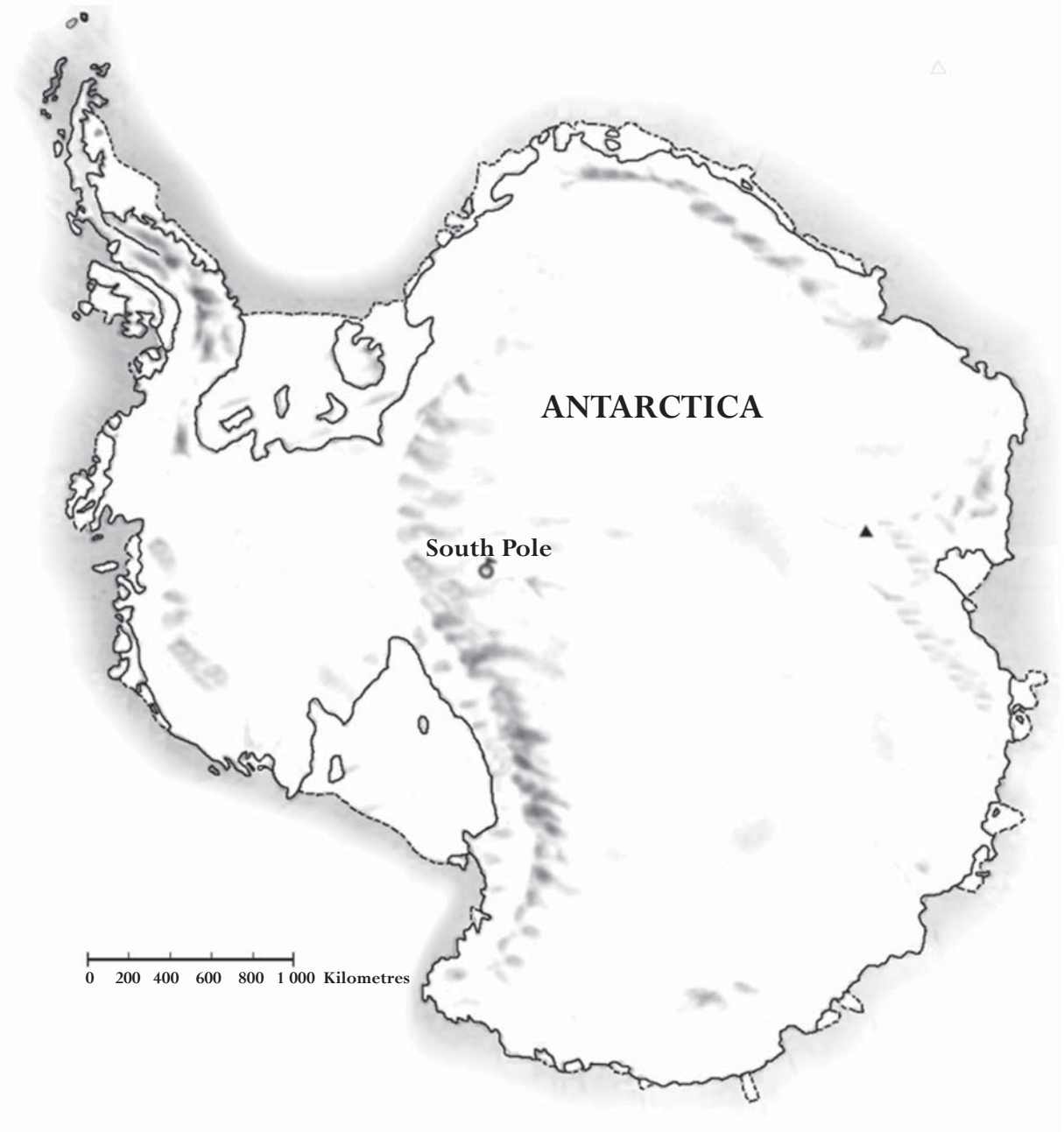
A field on a map looks like the figure below. (Scale 1:5000).

Estimate the actual area of the field.



CONTINENT AREA

Estimate the area of Antarctica using the map scale shown.



Mathematisation

Wherever possible, items should require students to carry out some mathematisation. Consider the following item.

Buy 2 get 1 free

TemTem biscuits are on special this week in both Supermarket A and Supermarket B.

Supermarket A advertised that if you buy 2 packets, you can get 1 packet free. Supermarket B advertised that the biscuits are 30% off the normal price.

(1) Which supermarket is offering a better deal, if the price before discount is the same at both supermarkets? Give reasons to support your answer.

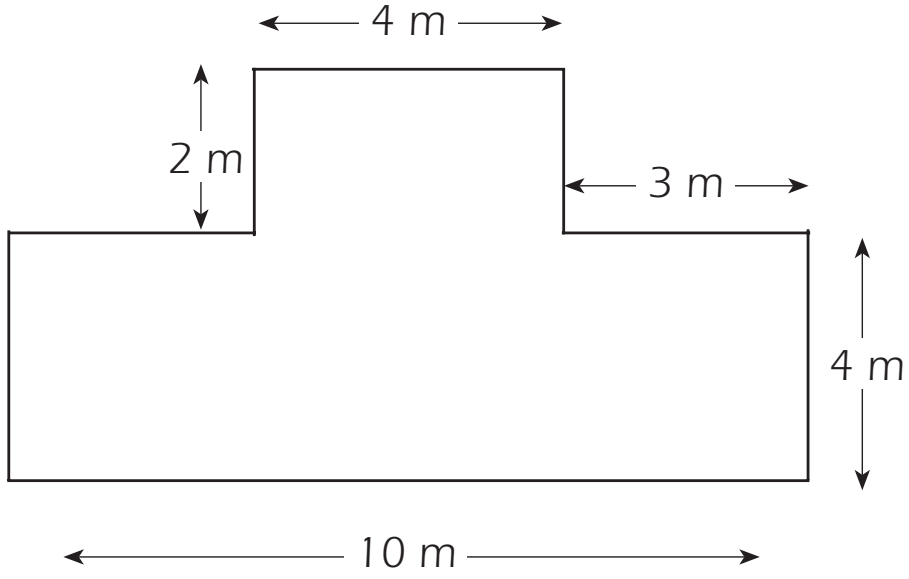
(2) Each packet of TemTem costs 7 zeds at Supermarket A. How much did Mrs Williams pay for 6 packets under the offer of "Buy 2 Get 1 Free"?

For this item, students have to work out the mathematical meaning of "Buy 2 Get 1 Free", a common advertising sign, but yet not a standard textbook application.

This following is another example on the degree of mathematisation given to the students. In the following, Carpenter 2 is a richer item than Carpenter 1, as it allows students to explore the properties of perimeter.

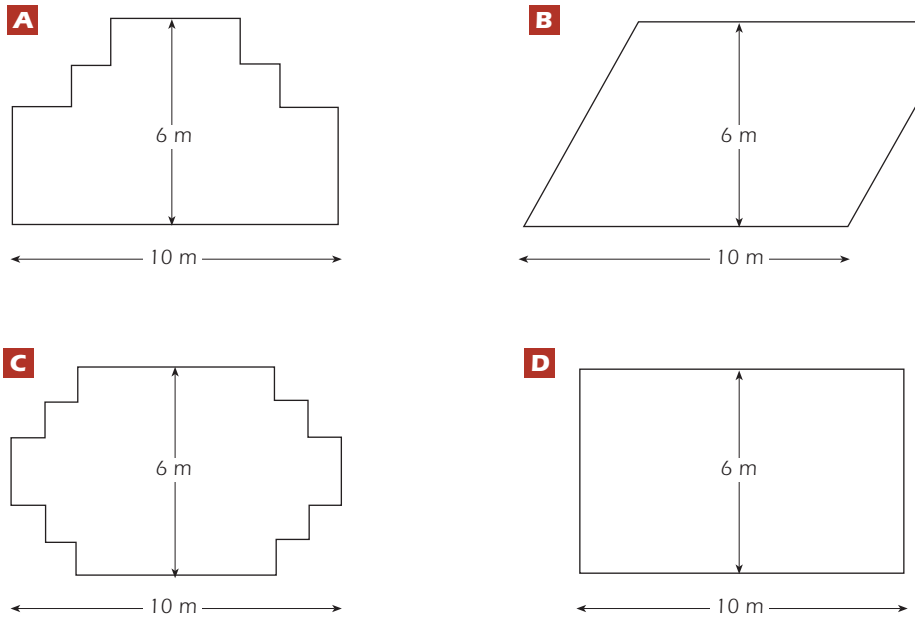
CARPENTER 1

A carpenter is building this framework. The measurements are given in metres. The framework encloses a garden bed. What length of wood will he have to buy?



CARPENTER 2

A carpenter has 32 metres of timber and wants to make a border around a garden bed. He is considering the following designs for the garden bed.



Which design(s) can be constructed with 32 metres of timber?

Mathematical context

The OECD PISA emphasis on authentic contexts does not preclude the inclusion of important and/or interesting mathematical contexts (sometimes these may be virtual contexts). The following item illustrates an interesting mathematical context.

The number 6 is called a perfect number because its factors (not including itself), 1, 2 and 3, add up to 6.

The next perfect number is 28, because its factors 1, 2, 4, 7, 14 add up to 28.

The next perfect number in the sequence is 496. Show that 496 is a perfect number.

(Note that the next two perfect numbers after 496 are 33550336 and 8589869056. Don't try to show that these are perfect numbers now!)

If the item simply said "List the factors of 496", then it is a standard textbook drill item. But by adding a mathematical context to this question, the item is made more interesting, and a purpose is added. That is, a list of the factors is requested in order to demonstrate a property of this number.

Data entry issues

Avoid using distracters that may be confused with the response labels or with score values.

How many cubes are painted blue?

A	0
B	1
C	2
D	3

Which one of the following is the corresponding track?

A	B
B	C
C	D
D	E

Avoid using scoring codes 7, 8 and 9 as these will be reserved as special score categories. If more than seven score categories are needed, then use double-digit codes.

Factors contributing to item difficulty

Introduction

Item difficulty is influenced by a number of factors that relate directly to the mathematical proficiency dimension that is being assessed through the test items. Items are developed and selected for inclusion in the PISA mathematics test instrument in such a way that these factors are consciously manipulated to enable assessment of the proficiency level of a wide range of students.

Other factors that influence item difficulty relate more to the clarity of the items – to aspects of item presentation and formulation. These factors, discussed earlier in this document, need to be identified and treated with the greatest of care in order to avoid the operation of factors that unintentionally affect item difficulty. Unintended factors should obviously be avoided – items should be made as easy to understand as possible. The incidence of unintended factors is related to the degree of thoroughness with which the items are written and developed.

The factors that will be intentionally varied to produce items that lie along the dimension of proficiency in mathematical literacy can be seen in the descriptions of proficiency levels for PISA mathematics. These factors include the following:

- The kind and degree of interpretation and reflection required. This includes the nature of interpretive demands arising from the problem context; the extent to which the mathematical demands of the problem are apparent or to which students must impose their own mathematical construction on the problem; and the extent to which insight, complex reasoning and generalisation are required.
- The kind and level of mathematical skill required, ranging from single-step problems requiring simple reproduction and computation skills through to multi-step problems involving more advanced mathematical knowledge, and complex decision making, cognitive processing, problem solving and modelling skills.

The following indicators should be useful in preparing items that test the required competencies, and that vary in difficulty.

Procedures and computation

Procedural and computational demands will range from application of single-step procedures, in familiar and well-defined problem settings that require the application of familiar computational processes, through to application of multi-step procedures in unfamiliar and ill-defined problem settings that require significant interpretation and mathematisation, and the selection and application of problem-solving skills and mathematical knowledge relevant to the problem at hand.

Representation

Task demands related to representation will vary from problems that are simply and fully stated with a clear representation of the problem already provided, through to problems where some decoding of representations may be needed, and where multiple representations may be required, with the student having to choose or devise useful representations of aspects of the problem, having to switch between different representations, and draw them together to work towards a solution.

Connection and integration

The simplest problems are those that involve single elements or the application of procedures without having to link different representations or different pieces of knowledge. Tasks become more demanding when they require students to draw connections between different mathematical techniques, or different areas of mathematical content, to combine and integrate information in order to tackle and solve problems, to choose and develop strategies, to choose mathematical tools, to use multiple methods or steps in the mathematisation and modelling process, and to interpret and reflect on a problem, on the meaning of a solution and on the validity of results.

Modelling and interpretation

The least demanding problems are those for which most or all of the modelling has already been done in the way the question is formulated. These tasks may require some interpretation, typically simple recognition of a familiar representation or process. Task demand is increased by requiring the student to interpret and structure the field or situation to be modelled, to make assumptions about which variables are important, to translate reality into mathematical structures, to interpret mathematical models in terms of reality, to work with and manipulate a mathematical model, and to validate the mode – this involves reflecting, analysing and offering a critique of a model and its results.

Reasoning and argumentation

Tasks demand reasoning and argumentation when the student is required to pose questions in order to formulate a problem, to distinguish between different kinds of statements (definitions, theorems, conjectures, hypotheses, examples, conditional assertions), to understand and handle the limits of given mathematical concepts, to follow a chain of mathematical reasoning, and to create mathematical arguments to explain, justify or communicate a result.

Insight and generalisation

Insight and generalisation are higher order processes generally associated with Class 3 competencies. Items demanding these processes require significant interpretation of complex material often in unfamiliar settings; posing of problems; creative linking of representations, processes or concepts with other knowledge; planning of solution strategies or significant modelling; and significant reflection on and presenting of mathematical outcomes, often with argument and generalisation.

Communication

The importance of communication is difficult to reflect in a pencil and paper test. PISA tasks impose communication demands when they involve significant reading and interpretation, and when they require students to explain or justify their results.

Section 3: Sample mathematics items

The following sample items are provided to illustrate the style of items wanted for PISA 2003. They were used in the PISA field trial in 1999. Some of these items were also reproduced in the OECD publication *Measuring Student Knowledge and Skills* (OECD, 2000).

Mathematics Sample Unit 1 – Pizzas

A pizzeria serves two round pizzas of the same thickness in different sizes. The smaller one has a diameter of 30 cm and costs 30 zeds. The larger one has a diameter of 40 cm and costs 40 zeds.

[© PRIM, Stockholm Institute of Education.]

Sample Question 1 (Open-constructed response)

Which pizza is better value for money? Show your reasoning.

Sample Question 1 scoring

Score 1: Gives general reasoning that surface area of pizza increases more rapidly than price of pizza to conclude that the larger pizza is better value

OR

Calculates the area and amount per zed for each pizza to conclude that the larger pizza is better value

Score 0: Other. Including a correct answer without correct reasoning.

Framework classification

This is a Competency Class 2 item. It is primarily associated with *change and relationships*. It is located in a personal situation.

Mathematics Sample Unit 2 – Seal's Sleep

A seal has to breathe even if it is asleep in the water. Martin observed a seal for one hour. At the start of his observation, the seal was at the surface and took a breath. It then dove to the bottom of the sea and started to sleep. From the bottom it slowly floated to the surface in 8 minutes and took a breath again. In three minutes it was back at the bottom of the sea again. Martin noticed that this whole process was a very regular one.

Sample Question 2 (Multiple choice)

After one hour the seal was

A At the bottom

B On its way up

C Breathing

D On its way down

Sample Question 2 scoring

Score 1: B

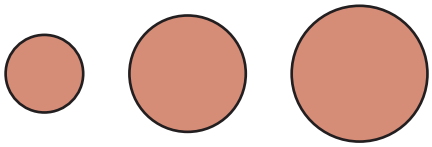
Score 0: Other

Framework classification

This is a Competency Class 2 item. It is primarily associated with *change and relationships.* It is located in a scientific situation.

Mathematics Sample Unit 3 – Coins

You are asked to design a new set of coins. All coins will be circular and coloured silver, but of different diameters.



Researchers have found out that an ideal coin system meets the following requirements:

- Diameters of coins should not be smaller than 15 mm and not be larger than 45 mm.
- Given a coin, the diameter of the next coin must be at least 30% larger.
- The minting machinery can only produce coins with diameters of a whole number of millimetres (e.g. 17 mm is allowed, 17.3 mm is not).

Sample Question 3 (Open-constructed response)

Design a set of coins that satisfy the above requirements. You should start with a 15 mm coin and your set should contain as many coins as possible.

Sample Question 3 scoring

Score 2: 15 – 20 – 26 – 34 – 45

Score 1: An answer that gives a set of coins that satisfy the three criteria, but not the set that contains as many coins as possible, for example: 15 – 21 – 29 – 39

OR

Answers that give the first four diameters correct, and the last one incorrect; or the first three diameters correct, with the last two incorrect

Score 0: Other incorrect responses.

Framework classification

This is a Competency Class 2 item. It is primarily associated with *change and relationships.* It is located in a occupational situation.

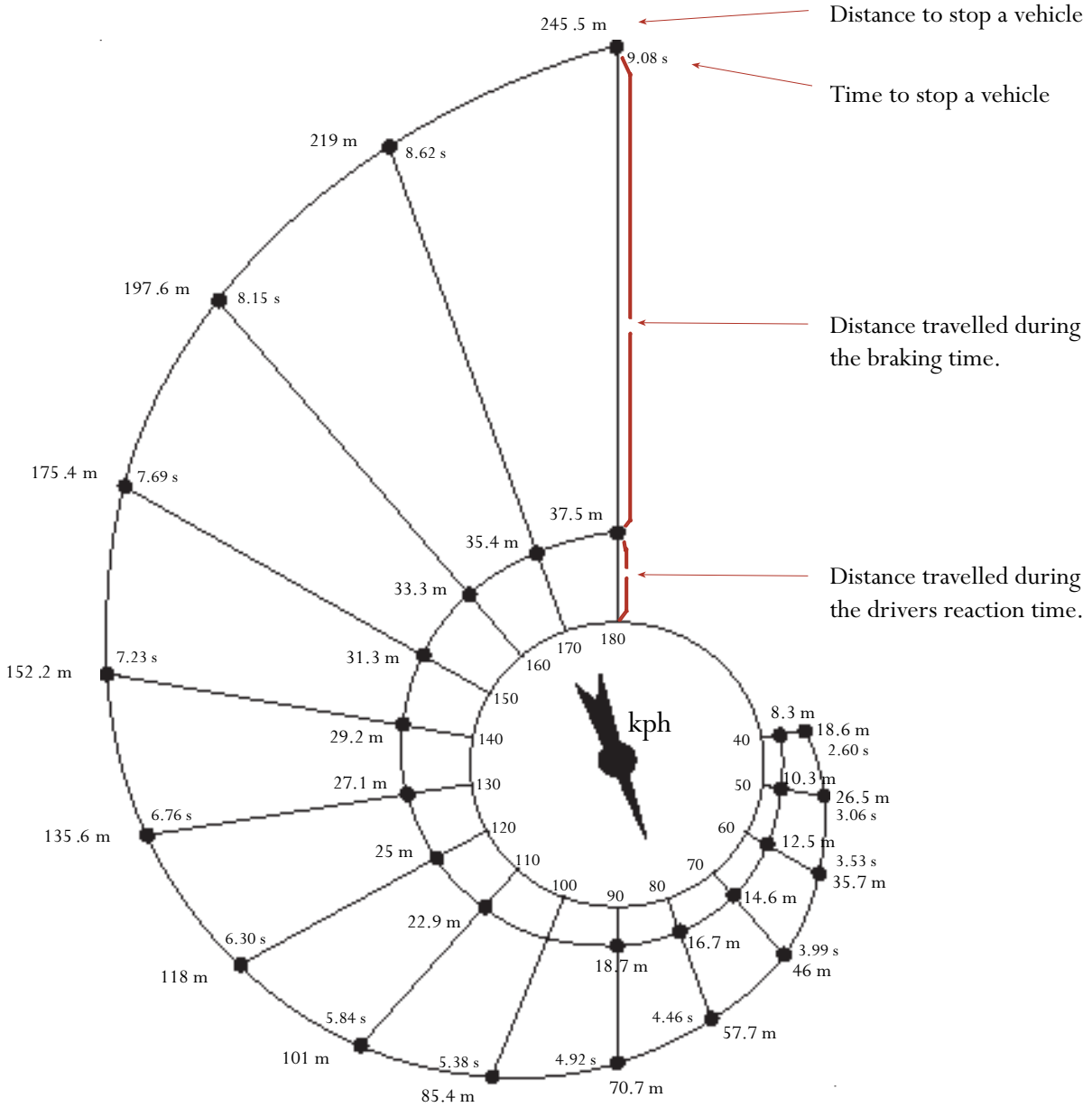
It is a partial credit item.

Mathematics Sample Unit 4 – Braking

The approximate distance to stop a moving vehicle is the sum of:

- The distance covered during the time the driver takes to begin to apply the brakes (reaction-time distance)
- The distance travelled while the brakes are applied (braking distance)

The snail diagram below gives the theoretical stopping distance for a vehicle in good braking conditions (a particularly alert driver, brakes and tyres in perfect condition, a dry road with a good surface) and how much the stopping distance depends on speed.



Sample Question 4 (Closed-constructed response)

If a vehicle is travelling at 110 kph, what distance does the vehicle travel during the driver's reaction time?

Sample Question 4 Scoring

Score 1: 22.9 metres (units not required)

Score 0: Other

Framework classification

This is a Competency Class 2 item. It is primarily associated with *change and relationships*. It is located in a personal situation.

Sample Question 5 (Closed-constructed response)

If a vehicle is travelling at 110 kph, what is the total distance travelled before the vehicle stops?

Sample Question 5 Scoring

Score 1: 101 metres (units not required)

Score 0: Other

Framework classification

This is a Competency Class 2 item. It is primarily associated with *change and relationships*. It is located in a personal situation.

Sample Question 6 (Closed-constructed response)

If a vehicle is travelling at 110 kph, what is the distance travelled while the brakes are being applied?

Sample Question 6 Scoring

Score 1: 78.1 metres (units not required)

Score 0: Other

Framework classification

This is a Competency Class 2 item. It is primarily associated with *change and relationships*. It is located in a personal situation.

OECD/PISA MATHEMATICS ITEM SUBMISSION FORM

Please include one completed copy of this form for each item submitted.

Name and country of sender: _____

Title of set of materials: _____

Author of set of materials: _____

Publication details: _____

Source in which material has appeared: _____

_____

Copyright permission:

☐ Being negotiated (details attached)

☐ Not required (source information attached)

☐ Obtained (copy of authorisation from copyright holder attached[2])

Competency Class:

☐ Class 1

☐ Class 2

☐ Class 3

Content aspect of these materials (problem domain):

☐ Quantity ☐ Space and shape

☐ Change and relationships ☐ Uncertainty

Situation of these materials:

☐ Personal ☐ Educational ☐ Occupational

☐ Public ☐ Scientific

☐ Other: _____

Item type:

☐ Multiple choice

☐ Closed-constructed response

☐ Open-constructed response

Appendix 5

ITEM REVIEW GUIDELINES

Notes for completing the item review forms

May 2002

Dear National Project Manager (NPM),

Introduction

The Consortium wishes to collect up-to-date item feedback information from National Centres to assist in the item selection process for the PISA 2003 main study.

Similar questions have been asked before, in the context of 'item bundle feedback'. However, a number of the field trial items have been substantially modified since the item bundle process, and your experience in preparing the items for use in the field trial will have given you additional and important insight into the items.

We now need to know your view of the field trial items, to assist in the next round of item selection. In addition, the BPC has made it very clear that it wants to be able to analyse PISA results partly on the basis of the considered item feedback given by National Centres. For this purpose, data of the highest possible quality and accuracy are required.

A set of Microsoft® Excel® spreadsheets has been prepared to assist you in recording your national experts' review of the items that were included in the 2002 field trial for PISA 2003.

It is important that you develop a well-considered response drawing upon appropriate expertise in your country, and that a single national response is returned to the consortium.

Separate spreadsheets have been provided for mathematics, science and problem solving. All field trial items are listed, and space is provided for ratings and comments related to a number of specific questions and issues.

Completing the spreadsheet

If your ratings/comments would be identical for the stimulus and all items within a unit, use the auto-fill feature of Excel to do this, or use copy and paste for the cells. It is important that you do this, because we will analyse the data at the item level, not the unit level.

Rating schemes

Five of the questions invite you to provide a rating (using the values 1 to 5). The general sense of each rating scheme is 1 = low; 2 = moderately low; 3 = medium level; 4 = moderately high; and 5 = high. The specific usage of the ratings for each question is described below.

Meaning of categories

The meanings of the categories used in the feedback spreadsheet are described in this section. The first four issues relate to relevance for 15-year-olds (curriculum relevance, PISA-relevance, interest, cultural relevance). Then there are questions on sensitivity issues, followed by technical issues from the field trial. Finally, you are asked for an overall priority rating.

Curricular match

How closely does the unit or item content correspond with material that would be dealt with in the typical school curriculum/curricula in your country, up to the stage(s) that 15-year-olds would have reached? Note that in each case, and for problem solving in particular, this match is not to the curriculum of a particular school subject, but to exposure to or familiarity with the relevant knowledge and skills from across the whole curriculum.

Use rating 1 (not in curriculum) through to rating 5 (standard curriculum material) to indicate how close the item is to school curriculum.

In the case where school curriculum varies within your country, use the ratings to indicate the percentage of students who would have covered the content area of the item. Use rating 1 (0 to 20 per cent); rating 2 (20 to 40 per cent); rating 3 (40 to 60 per cent); rating 4 (60 to 80 per cent); and rating 5 (80 to 100 per cent) to express your estimate of the proportions.

Match to PISA objectives and framework

How relevant is the task for the students in preparing for life and other broad intentions defined in the OECD PISA frameworks? We refer to skills that are needed in many facets of one's life, such as for work, leisure, and participation in the society. These are not only basic life skills in everyday life. Use rating 1 for "not relevant to PISA objectives", through to rating 5 for "highly relevant to PISA objectives".

Interest level

How interesting is the task for the students? Here we refer to a variety of motivational aspects of the task: Would students find the problem stimulating? Would the students feel rewarded after solving the problem (the "a-ha" feeling)? Does the item have an interesting context? Does the item have an interesting diagram or graph? Is the solution unexpected? Would students relate the task to their personal experience? Use rating 1 for "not interesting" through to rating 5 for "extremely interesting".

Cultural relevance

How relevant is the item for students in your country from a cultural point of view? Is it important in the national context for students to be able to complete such a task? Use rating 1 for "not relevant" through to rating 5 for "extremely relevant".

Sensitivity concerns – Yes/No and Comments

Are there any concerns for the items regarding cultural sensitivities? For example, is the topic acceptable for 15-year-old students in your country? Answer "yes" or "no" in the sensitivity concerns column, for all items, and for those items where you answer "yes", specify your concern(s) in the comments column.

Note that in an international test, we cannot possibly have all items with settings that are familiar to students in every country. Sensitivity concerns refer to problem settings that are not acceptable in the country, for reasons such as religion or culture. For example, the promotion of the legalisation of drug use may not be acceptable as the context for an item. Sensitivity concerns are not about particular contexts that do not exist in your country, unless the familiarity of the context is likely to differentially affect item difficulty. A separate rating opportunity is provided for cultural relevance that covers these more general issues.

Translation or adaptation problems

When preparing the material for the field trial, were there any significant translation problems with the wording of the items? This is not only about how easy it is to translate the item into your language. It is about whether you were able to maintain the same level of item difficulty after you translated the item for the field trial. Please write a brief description of any translation problems that occurred.

Coding/marking problems

When coding student responses to the field trial items, did your markers experience any significant coding problems with the items? Please write a brief description of any marking and coding problems that occurred.

Other comments

If you have any comments that are not covered under the previous headings, please free feel to add any comments here. By the way, we are not only looking for comments about potential problems. If you particularly like an item, we would like to hear about it too.

Priority for inclusion

The final rating should be used to give an on-balance judgement about each item, and its value in an international test. Your judgement should combine all the information you have about the item (curricular relevance, interest, relevance to the PISA framework, cultural relevance, sensitivity issues, and technical matters), to indicate whether or not you think the item should be included in the international selection. Use the ratings '1' to show you assign a low priority to having that item included in the main survey; '2' to show moderately low priority; '3' to show a medium level priority; '4' to show moderately high priority; and '5' to show that the unit/item is one you regard as having highest priority for inclusion.

Completing the spreadsheet electronically

The consortium requests that you complete the spreadsheet electronically. The cells have been formatted to allow text to wrap and row height to grow if you type in a lot of text.

Please e-mail the completed spreadsheet to the following address by no later than Friday, 23 August 2002: *pisa@acer.edu.au*

Appendix 6

ISCED ADAPTATIONS FOR PARTNER COUNTRIES

This appendix lists adaptations to the International Standard Classification of Education categories for those countries not included in the Manual for ISCED-97 (OECD, 1999). Adaptations for the 29 OECD countries follow the classification elaborated in ISCED-97. Recent changes in the educational system in some countries are reflected in the adaptations for these countries.

Brazil

	National programme names	English programme names
ISCED 1	Ensino fundamental de 1a. a 4a. série	Primary education from 1st to 4th grade
ISCED 2	Ensino fundamental de 5a. a 8a. série	Lower secondary education from 5th to 8th grade
ISCED 3B,C	N/A	N/A
ISCED 3A	Ensino médio (regular ou técnico)	Upper secondary education – general or technical
ISCED 4	N/A	N/A
ISCED 5B	Curso técnico superior	Technical graduation
ISCED 5A	Curso superior	Graduation
	Pós-graduação	Post-graduation

Hong Kong-China

	National programme names	English programme names
ISCED 1	小學	Primary school
ISCED 2	中三	Form 3 /Grade 9
ISCED 3B,C	中五文法課程	Form 5/Grade 11 grammar or international programme
	中五職先或技術課程	Form 5 /Grade 11 pre-vocational or technical programme
ISCED 3A	中六-中七	Form 6-7 /Grade 12-13
ISCED 4	中五持續進修課程(例如：毅進計劃、副學士先修課程)	Post-secondary 5 continuing education (*e.g.* project springboard or pre-associate degree)
ISCED 5B	專上學院課程（非學位課程，例如：高級文憑）	College (non-degree, *e.g.* higher diploma)
ISCED 5A	學士課程或以上	Bachelor or above

Indonesia

	National programme names	English programme names
ISCED 1	Tamat SD/ MI	Primary school
ISCED 2	Tamat SLTP/ MTs	Junior secondary, general
ISCED 3B,C	Tamat SMK, LPK (kursus)	Senior secondary, technical/vocational
ISCED 3A	Tamat SMU/ MA	Senior secondary, general
ISCED 4	Diploma 1 atau 2	Diploma 1 and 2
ISCED 5B	Politeknik/Akademi/D3 atau D4	Polytechnical, academy, diploma 3 and 4
ISCED 5A	S1, S2, atau S3	Specialist programmes 1, 2 and 3 (including masters' and doctoral degree)

Latvia

	National programme names	English programme names
ISCED 1	Sākumskolu	Primary school
ISCED 2	Pamatskolu	Basic school
ISCED 3B,C	Arodskolu vai tehnikumu	Vocational or technical school
ISCED 3A	Vidusskolu	Secondary school
ISCED 4	Izglītība pēc vidusskolas, bet ne augstskolā	Education after secondary school (but not in university)
ISCED 5B	Augstskolas profesionālo studiju programma (instituta ieguta augstaka izglītība)	Professional program (higher education at institute level)
ISCED 5A	Augstskolas bakalaura vai magistra studiju programma (universitate ieguta augstaka izglitiba)	Bachelor or master program (higher education in university)

Liechtenstein

Same adaptations as for Switzerland

Macao-China

Same adaptations as for Hong Kong-China

Russia

	National programme names	English programme names
ISCED 1	Начальное общее образования (1-4 классы)	Primary general education (grades 1-4)
ISCED 2	Основное общее образование (5-9 классы)	Basic general education (grades 5-9)
ISCED 3B,C	Начальное профессиональное образование (например, профессиональное училище) Среднее профессиональное образование (например, техникум)	Initial professional education (e.g. professional school) Secondary professional education (e.g. technicum)
ISCED 3A	Среднее общее образование (10-11 классы)	Secondary general education (grades 10-11)
ISCED 4	После окончания средней школы закончила любые курсы продолжительностью от 6 месяцев до 2-х лет.	After graduating the secondary school finished any professional cources from six months to two years.
ISCED 5A	Высшее образование (закончила институт, университет, академию или аспирантуру)	Higher education (finished institute, university, academy or doctorate courses)

Slovak Republic

	National programme names	English programme names
ISCED 1	stupeň základnej školy	ISCED level 1
ISCED 2	základnú školu (2. stupeň)	Second level of basic school
ISCED 3B,C	odborné učilište bez maturity	Vocational college
ISCED 3A	gymnázium, strednú odbornú školu alebo učilište s maturitou	Secondary school, secondary college or technical college
ISCED 4	nejaké nadstavbové štúdium	Post-secondary qualification studies
ISCED 5B	bakalárske štúdium, prípadne vyššiu odbornú školu	Bachelor, higher professional studies
ISCED 5A	vysokoškolské štúdium, možno aj doktorát	University studies (master's and doctoral degrees)

Thailand

	National programme names	English programme names
ISCED 1	ระดับประถมศึกษ	Primary level
ISCED 2	ระดับมัธยมศึกษาตอนต้น	Lower secondary level
ISCED 3B,C	ระดับประกาศนียบัตรวิชาชีพ(ปวช)	Vocational school or technical college
ISCED 3A	ระดับมัธยมศึกษาตอนปลายสายสามัญ	Upper secondary level
ISCED 4	อนุปริญญา หรือ ประกาศนียบัตรวิชาชีพชั้นสูง (ปวส)	Diploma or higher vocational certificate/technical certificate
ISCED 5A	ปริญญาตรีหรือสูงกว่า	Bachelor's degree or higher

Tunisia

	National programme names	French programme names
ISCED 1	الابتـــــدائي التعليـــــم مســـتوى	Niveau primaire
ISCED 2	الإعـدادي التعليـــــم مســـتوى	2ème cycle de l'enseignement de base (collège)
ISCED 3B,C	تكـوين مهـني بعـد المدرســة الإعداديــة	Formation professionnelle après le collège
ISCED 3A	الباكلوريـــــا مســـتوى	Niveau du baccalauréat
ISCED 4	شـهادة) ســـــنتين + الباكلوريـــــا شـهادة ...سـامي تقـــني)	Le baccalauréat + 2 années d'études (technicien supérieur…)
ISCED 5B	ختـم شـهادة) الباكالوريـــــا بعـد ســـنتان شـهادة ، العـــالي التعليم من الأولـى المرحلـة لتكـــــوين الأعلـى المعهـد دروس ختـم ...المعلّمين)	Bac + 2 (DUEL, diplôme d'instituteur, etc…)
ISCED 5A	المعمّقة الدّراسات) فـوق ما أو الأستاذيّة شـهادة الـــدكتوراه أو)	Maîtrise et plus (DEA, doctorat)

Uruguay

	National programme names	English programme names
ISCED 1	Primaria	Primary school
ISCED 2	Cursos Básicos, Formación Profesional Básica o Ciclo Básico en UTU. Ciclo Básico de Secundaria	Basic courses, professional basic education, basic cycle UTU, basic cycle of secondary
ISCED 3B,C	Bachillerato Tecnológico o Formación Profesional Superior en UTU	Technical diploma or higher professional education
ISCED 3A	Bachillerato Diversificado en liceos públicos o privados	Diploma from public or private high schools
ISCED 5B	Un título terciario intermedio corres-pondiente a carreras cortas o técnicas tales como auxiliar de enfermería, auxiliar contable, técnico en informática, etc	Post-secondary technical careers or short university careers.
ISCED 5A	Un título universitario, docente o de postgrado	University degree, teaching degree or post-graduate

Serbia (Serbia and Montenegro)

	National programme names	English programme names
ISCED 1	Cetiri razreda osnovne skole	Lower primary school
ISCED 2	Osnovnu skolu	Primary school (lower secondary)
ISCED 3B,C	Srednju strucnu skolu, trogodisnju ili cetvorogodisnju	Vocational secondary school, lasting three or four years
ISCED 3A	Gimnaziju	Gymnasium
ISCED 4	Visu skolu	Higher school
ISCED 5B	Fakultet	Faculty
ISCED 5A	Magistraturu ili doktorat	Masters' and doctoral degrees

Appendix 7

FICTITIOUS EXAMPLE OF A STUDY PROGRAMME TABLE (SPT)

Study programme table

Country: **<country name>**

NPM: **<name>**

Date: **<date>**

G	General
V	Vocational
P	Pre-vocational
First grade	Lowest possible grade
Last grade	Highest possible grade

1 Programme No.	2 Programme Code in STF	3 National name of programme	4 English/French programme description	5 First grade	6 Last grade	7 ISCED level	8 ISCED designation	9 Orientation	10 Comments
1	1	Secundaria básica	Lower secondary	7	9	2	A	G	
2	1	Secundaria superior	Upper secondary	10	12	3	A	G	
3	2	Secundaria técnica	Technical upper secondary	1	3	3	B	P	Grades +9
4	3	Educación profesional	Vocational education	97	97	3	C	V	Ungraded programme

Appendix 8

FICTITIOUS EXAMPLE OF THE QUESTIONNAIRE ADAPTATION SPREADSHEET (QAS)

1	2	3	4	5	6	7	8	9	10	11
Q_Int	**English version**	**N_Int**	**N_XXX**	**Lab_Int**	**Lab_XXX**	**Q_XXX**	**National version**	**Translation of the national version**	**Justification for proposed changes**	***Queries / approval***
Q12	**Does your mother have any of the following qualifications?**					**11**				
a	<ISCED 5A or 6>			ST12Q01	**ST12N01**	**a**	Doctorado	PhD	ISCED 6	
	AS PREVIOUS				**ST12N02**	**b**	Bachiller	Bachelor's Degree	ISCED 5A	
	AS PREVIOUS				**ST12N03**	**c**	Diploma de colegio técnico	Technical College Diploma	ISCED 5A	
b	<ISCED 5B>			ST12Q02		**d**	Certificado profesional	Professional certificate	ISCED 5B	
c	<ISCED 4>			ST12Q03			**DELETED**		This level does not exist in our country.	
	Tick	1	1							
	No tick	2	2							

Appendix 9

SUMMARY OF QUALITY MONITORING OUTCOMES

Not all information gathered by PISA quality monitors (PQM), or national centre quality monitors (NCQM), is used in this report. Items considered relevant to data quality and to the conduct of the test sessions are included in this appendix.

PISA quality monitors

A total of 109 PQMs submitted 645 reports on the conduct of testing sessions in all countries participating in PISA. Each PQM also submitted a report detailing their general observations of the testing procedures in the country they were monitoring.

Conditions for the test and questionnaire

The data from the PQM reports suggested that the preparations for the PISA test sessions went well and that test room conditions were adequate. Where problems with conditions were identified, the PQM comments indicate that these were due to normal school activities such as noise from other classes, noise from changing of classes, and disruptions caused by school announcements.

Figure A9.1
PISA quality monitors' comments on the test conditions

Was the test area suitable for the assessment by providing:

	Yes		No		Missing	
A reasonable amount of space	619	(96.0%)	23	(3.6%)	3	(0.5%)
Sufficient light	640	(99.2%)	3	(0.5%)	2	(0.3%)
A quite testing environment	606	(94.0%)	29	(4.5%)	10	(1.6%)
Isolation from school distractions	591	(91.6%)	39	(6.0%)	15	(2.3%)
A comfortable temperature to work	628	(97.4%)	13	(2.0%)	4	(0.6%)

Were there any general disruptions to the session that lasted for more than one minute (e.g. alarms, announcements, changing of classes, etc.?)

Yes	72	(11.2%)
No	571	(88.5%)
Missing	2	(0.3%)

Conducting the testing sessions

A detailed script for conducting the PISA test sessions was provided through the PISA Test administrator manual. While some variation in the script was permitted to suit local circumstances, test administrators were reminded to follow the script as closely as feasible. In each monitored testing session, the PQM recorded the test administrator's adherence, or lack thereof, to the script. The instances where the test administrator followed the instructions verbatim was high (82.1%).

The PQM reports indicated that PISA test administration procedures were followed in most cases. A failure to properly record student attendances was identified in six (0.9%) cases. A failure to properly record the test timing information was identified in seven (1.1%) cases. PQMs identified 47 (7.3%) cases where students were admitted to the testing room after the testing session had begun. This represents a violation of PISA procedures and may be due to insufficient emphasis on this aspect during test administrator training. Defective booklets were detected 50 (7.8%) times and these were replaced correctly by the test administrator in the majority of cases.

Students

The data suggest that a high proportion of students who participated in the PISA tests and questionnaire did so in a positive spirit. While attempts at cheating were reported in 26 (4.0%) cases, it is highly unlikely that these occurrences would have any influence on results as students had different test booklets.

Figure A9.2
PISA quality monitors' comments on test administrator adherence

Did the test administrator read the script exactly as it is written?

Yes	5174	(81.2%)	
No	1198	(18.8%)	If no, did the test administrator make:
Minor additions	620	(9.7%)	
Major additions	56	(0.9%)	
Minor deletions	296	(4.6%)	
Major deletions	120	(1.9%)	
Missing	106	(1.7%)	

Where major additions or deletions were identified they generally arose when the test administrator paraphrased instructions.

Figure A9.3
PISA quality monitors' comments on the students

Generally, were the students orderly and co-operative?

Yes	619	(96.0%)
No	24	(3.7%)
Missing	2	(0.3%)

Did any students refuse to participate in the assessment after the session had begun?

Yes	25	(3.9%)
No	611	(94.7%)
Missing	9	(1.4%)

Was there any evidence of students cheating during the assessment session?

Yes	26	(4.0%)
No	615	(95.3%)
Missing	4	(0.6%)

Most of the evidence to do with cheating concerned students exchanging booklets, asking each other questions, and showing each other their answers. Some students checked with their neighbour to see if they had the same test booklet. Most incidences reported by the PQMs seem to have been relatively minor.

Interview with school co-ordinator

PQMs conducted an interview with the PISA school co-ordinators to identify problems experienced by schools in conducting PISA. The data from the interview suggest that in the majority of cases schools did not have difficulty in securing the support of parents and teachers.

Figure A9.4
School co-ordinators' comments on conducting PISA

Were there any difficulties in securing parental permission

Not difficult	494	(76.6%)
Some difficulties	42	(6.5%)
Very difficult	5	(0.8%)
Not applicable	96	(14.9%)
Missing	8	(1.2%)

Were there any difficulties in obtaining support from teachers

Not difficult	553	(85.7%)
Some difficulties	44	(6.8%)
Very difficult	4	(0.6%)
Not applicable	39	(6.0%)
Missing	5	(0.8%)

Were there any difficulties in organising a suitable testing area

Not difficult	545	(84.5%)
Some difficulties	64	(9.9%)
Very difficult	11	(1.7%)
Not applicable	23	(3.6%)
Missing	2	(0.3%)

For your school, is there another time in the school year more suitable for an assessment such as PISA?

Yes	254	(39.4%)
No	380	(58.9%)
Missing	11	(1.7%)

Did you have any problems making exclusions after you received the list of sampled students?

Yes	37	(5.7%)
No	602	(93.3%)
Missing	6	(0.9%)

In the majority of cases where the school co-ordinator identified a more suitable testing time, the reason provided related to local school circumstances such as the number of student illnesses, school camps and other local functions.

In most cases where the school co-ordinator reported problems with exclusions the problem related to students with special circumstances. Examples include students with attention deficit hyperactivity disorder, students with chronic absenteeism, and a student who was a professional athlete. Other problems related to errors in students listed on the student tracking form.

Security of materials

PQMs reported on the overall security of materials in their general observations report completed after their school visits. In only one case did a PQM observe a potential breach of security. In this single case the PQM observed that a box with PISA

materials had been opened at a school before the test administrator had arrived. However, in this single case the PQM was confident that the material had not been abused. No other actual or potential security breaches were identified by the PQMs.

National centre quality monitors

NCQMs interviewed all national project managers (NPM) using a standard schedule of questions. Most interviews lasted about two hours.

General organisation of PISA

The consortium regularly sought national feedback during the development of test and questionnaire instruments. To ensure that feedback represented a range of viewpoints within a country, NPMs were encouraged to establish committees to provide advice on PISA matters. A majority of NPMs were able to follow this advice. However, the fact that at least eight countries (20%) did not have national committees suggests that this matter should be addressed in future cycles so that the validity of PISA can be maintained.

Figure A9.5
National Project Manager comments on the general organisation of PISA

How would you categorise the organisation around which PISA is organised in this country?

A university	6	(15.0%)
A government body	21	(52.5%)
A consortium of groups	3	(7.5%)
A not-for-profit organisation	3	(7.5%)
Other	7	(17.5%)

Is the NPM and BPC member the same person in this country?

Yes	10	(25.0%)
No	30	(75.0%)

Have any national committees been established to support or advise the national centre on PISA matters?

Yes	30	(75%)
No	8	(20%)
Missing	2	(5%)

Test administrators

PISA has established clear criteria for the selection of test administrators, most NPMs found these relatively easy to comply with. Among the difficulties encountered by NPMs was difficulty in finding test administrators for certain regions. Also, in the case of countries that took a census for PISA it was not possible to find teachers that were not a member of staff at a sampled school.

Figure A9.6
National Project Manager comments on selecting test administrators

Was it easy to comply with these criteria?

Yes, easy	26	(65.0%)
No, not easy	13	(32.5%)
Missing	1	(2.5%)

Who will be the test administrators for the main study?

National centre staff	5	(12.5%)
Regional or district staff	4	(10.0%)
External contractor	6	(15.0%)
Teacher but not of sampled student	3	(7.5%)
Teacher of sampled student but not in a PISA domain	4	(10.0%)
Other school staff	2	(5.0%)
Other, specify	12	(30.0%)
Missing	4	(10.0%)

Note: In most cases where NPMs indicated "other" the arrangements were a hybrid that included two or more of the options provided in the question.

Figure A9.7
National Project Manager comments on coders

Coders

- *must have a good understanding of mid-secondary level mathematics and science OR <test language>;*
- *understand secondary level students and the ways that students at this level express themselves.*

Do you anticipate having any problems, or did you have problems, recruiting the coders who met these criteria?

Yes	15	(37.5%)
No	24	(60%)
Missing	1	(2.5%)

From where were, the coders recruited? (Or, from where will the coders be recruited?)

National centre staff	3	(7.5%)
Teachers/professional educators	15	(37.5%)
University students	10	(25.0%)
Other, specify	8	(20.0%)
Missing	4	(10.0%)

Coders

PISA has established guidelines for the selection of manual coders. At the time of the NCQM interview, a large minority of 15 NPMs (37.5%) indicated that they would have problems meeting these criteria. The major problems cited were the availability of qualified staff and that the coding activity would occur at a time when most teachers had other obligations. Three NPMs did not have representatives attend the international coder meeting in February 2003, and for these three countries the consortium initiated corrective action in the form of supplementary training opportunities. Most countries used table leaders to check the quality of the coding process.

In most cases where NPMs indicated 'other' the arrangements were a hybrid that included two or more of the options provided in the question. Two NPMs indicated that they would employ trainee teachers.

Conclusion

In general, the quality monitoring reports suggest a strong organisational base within countries for the conduct of PISA. The NCQM reports indicate that, in the main, NPMs and National Centre staff had a good understanding of the operational aspects of PISA. The PQM reports indicate that the PISA test and questionnaire sessions were conducted in a manner that was largely consistent with the documented procedures in the PISA operations manuals.

Appendix 10

CONTRAST CODING FOR PISA 2003 CONDITIONING VARIABLES

Conditioning variables	Variable name(s)	Variable coding	Contrast coding
What <grade> are you in? Q1a	ST01Q01	7	7 0
		8	8 0
		9	9 0
		10	10 0
		11	11 0
		12	12 0
		13	13 0
		Missing	(mean)1
Gender Q3	ST03Q01	1 <Female>	-1
		2 <Male>	1
		Missing	0
What is your mother currently doing? Q5	ST05Q01	1 <Working full-time <for pay>	0000
		2 <Working part-time <for pay>	1000
		3 <Not working, but looking for a job>	0100
		4 <Other (*e.g.* home duties, retired) >	0010
		Missing	0001
What is your father currently doing? Q6	ST06Q01	1 <Working full-time <for pay> >	0000
		2 <Working part-time <for pay> >	1000
		3 <Not working, but looking for a job>	0100
		4 <Other (*e.g.* home duties, retired) >	0010
		Missing	0001
Age when arrived to the country of test Q15b	ST15Q04	Value (decimal)	(copy)0
		Missing	(mean)1
Language at home Q16	ST16Q01	1 '<Test language>'	0000
		2 '<Other national language>'	1000
		3 '<Other national dialects>'	0100
		4 '<Other languages>'	0010
		Missing	0001
Possessions own room Q17b	ST17Q02	1 <Tick>	1
		2 <Not Tick>	0
Possessions dishwasher Q17m	ST17Q13	1 <Tick>	1
		2 <Not Tick>	0
How many books at home Q19	ST19Q01	1 <0-10 books>	100000
		2 <11-25 books>	010000
		3 <26-100 books>	001000
		4 <101-200 books>	000000
		5 <201-500 books>	000100
		6 <More than 500 books>	000010
		Missing	000001
Attend <ISCED 0> Q20	ST20Q01	1 <Not Tick>	000
		2 <Yes, one year or less>	100
		3 <Yes, more than one year>	010
		Missing	001
<ISCED 1>Years Q21	ST21Q01	value (decimal)	(copy)0
		Missing	(mean)1
Repeat <ISCED 1> Q22a	ST22Q01	1 <No, never>	000
		2 <Yes, once>	100
		3 <Yes, twice or more>	010
		Missing	001
Repeat <ISCED 2> Q22b	ST22Q02	1 <No, never>	000
		2 <Yes, once>	100
		3 <Yes, twice or more>	010

Conditioning variables	Variable name(s)	Variable coding	Contrast coding
		Missing	001
Repeat <ISCED 3> Q22c	ST22Q03	1 <No, never>	000
		2 <Yes, once>	100
		3 <Yes, twice or more>	010
		Missing	001
Attend local Q25a	ST25Q01	1 <Tick>	10
		2 <Not Tick>	00
		Missing	01
Attend better Q25b	ST25Q02	1 <Tick>	10
		2 <Not Tick>	00
		Missing	01
Attend specific program Q25c	ST25Q03	1 <Tick>	10
		2 <Not Tick>	00
		Missing	01
Attend religious Q25d	ST25Q04	1 <Tick>	10
		2 <Not Tick>	00
		Missing	01
Attend family Q25e	ST25Q05	1 <Tick>	10
		2 <Not Tick>	00
		Missing	01
Attend Other Q25f	ST25Q06	1 <Tick>	10
		2 <Not Tick>	00
		Missing	01
Late for school Q28	ST28Q01	1 <None >	0000
		2 <1 or 2 times>	1000
		3 <3 or 4 times>	0100
		4 <5 or more times>	0010
		Missing	0001
Hours all homework Q29a	ST29Q01	Value (decimal)	(copy)0
		Missing	(mean)1
Hours all <Remedial> Q29b	ST29Q02	Value (decimal)	(copy)0
		Missing	(mean)1
Hours all <Enrichment> Q29c	ST29Q03	Value (decimal)	(copy)0
		Missing	(mean)1
Hours all tutor Q29d	ST29Q04	Value (decimal)	(copy)0
		Missing	(mean)1
Hours all <out-of-school> Q29e	ST29Q05	Value (decimal)	(copy)0
		Missing	(mean)1
Hours all other study Q29f	ST29Q06	Value (decimal)	(copy)0
		Missing	(mean)1
Hours mathematics homework Q33a	ST33Q01	Value (decimal)	(copy)0
		Missing	(mean)1
Hours mathematics <Remedial> Q33b	ST33Q02	Value (decimal)	(copy)0
		Missing	(mean)1
Hours mathematics <Enrichment> Q33c	ST33Q03	Value (decimal)	(copy)0
		Missing	(mean)1
Hours mathematics tutor Q33d	ST33Q04	Value (decimal)	(copy)0
		Missing	(mean)1
Hours mathematics <out-of-school> Q33e	ST33Q05	Value (decimal)	(copy)0
		Missing	(mean)1
Hours mathematics other Q33f	ST33Q06	Value (decimal)	(copy)0
		Missing	(mean)1
Students in mathematics Q36	ST36Q01	0	0 0

Conditioning variables	Variable name(s)	Variable coding	Contrast coding
		1	1 0
		2	2 0
		…	…
		90	900 0
		Missing	(mean)1
Lesson book work Q38d	ST38Q04	1 <Every lesson >	1 0
		2 <Most lessons >	2 0
		3 <Some lessons >	3 0
		4 <Never or hardly ever >	4 0
		Missing	(mean)1
Miss two or more months of <ISCED 1> EC1	EC01Q01	1 <No, never>	000
		2 <Yes, once>	100
		3 <Yes, twice or more>	010
		Missing	001
Miss two or more months of <ISCED 2> EC2	EC02Q01	1 <No, never>	000
		2 <Yes, once>	100
		3 <Yes, twice or more>	010
		Missing	001
Change schools while in <ISCED 1> EC3	EC03Q01	1 <No, never>	000
		2 <Yes, once>	100
		3 <Yes, twice or more>	010
		Missing	001
Change schools while in <ISCED 2> EC4	EC04Q01	1 <No, never>	000
		2 <Yes, once>	100
		3 <Yes, twice or more>	010
		Missing	001
Change <study programme> since <Grade X> EC5	EC05Q01	1 <Yes>	10
		2 <No>	00
		Missing	01
Type <mathematics class> EC6	EC06Q01	1 <high level>	100
		2 <medium level>	000
		3 <basic level>	010
		Missing	001
Mark in <mathematics> EC7a	EC07Q01	Value (decimal)	(copy)0
		Missing	(mean)1
Pass mark in mathematics EC7b	EC07Q02	1 <At or above <pass mark>>	10
		2 <Below <pass mark>>	00
		Missing	01
Mark in Maths in percentages EC7c	EC07Q03	value (decimal)	(copy)0
		missing	(mean)1
Available at home IC1a	IC01Q01	1 <Yes>	10
		2 <No>	00
		Missing	01
Available at school IC1b	IC01Q02	1 <Yes>	10
		2 <No>	00
		Missing	01
Available at other places IC1c	IC01Q03	1 <Yes>	10
		2 <No>	00
		Missing	01
Used computer IC2	IC02Q01	1 <Yes>	10
		2 <No>	00
		Missing	01

Conditioning variables	Variable name(s)	Variable coding	Contrast coding
How long using computers IC3	IC03Q01	1 <Less than 1 year >	0000
		2 <1 to 3 years >	1000
		3 <3 to 5 years >	0100
		4 <More than 5 years >	0010
		Missing	0001
Use often at home IC4a	IC04Q01	1 <Almost every day >	1 0
		2 <A few times each week >	2 0
		3 <Between once per week and once per month >	3 0
		4 <Less than once per month>	4 0
		5 <Never >	5 0
		Missing	(mean) 1
Use often at school IC4b	IC04Q02	1 <Almost every day >	1 0
		2 <A few times each week >	2 0
		3 <Between once per week and once per month >	3 0
		4 <Less than once per month>	4 0
		5 <Never >	5 0
		Missing	(mean) 1
Use often at other places IC4c	IC04Q03	1 <Almost every day >	1 0
		2 <A few times each week >	2 0
		3 <Between once per week and once per month >	3 0
		4 <Less than once per month>	4 0
		5 <Never >	5 0
		Missing	(mean) 1
Learn Computer IC8	IC08Q01	1 < My school >	00000
		2 < My friends >	10000
		3 < My family >	01000
		4 < Taught myself >	00100
		5 < Others >	00010
		Missing	00001
Learn Internet IC9	IC09Q01	1 < Don't know how to use >	000000
		2 < My school >	100000
		3 < My friends >	010000
		4 < My family >	001000
		5 < Taught myself >	000100
		6 < Others >	000010
		Missing	000001
Student questionnaire - composite variables			
Student age	AGE	Value (decimal)	(copy) 0
		Missing	(mean) 1
Country of birth	IMMIG	1 <Native students>	000
		2 <First-Generation students>	100
		3 <Non-native students>	010
		Missing	001
Mother's highest qualifications?	MISCED	0 <none >	1000000
		1 <ISCED1 >	0100000
		2 <ISCED2 >	0010000

Conditioning variables	Variable name(s)	Variable coding	Contrast coding
		3 <ISCED3B,C>	0001000
		4 <ISCED3A >	0000000
		5 <ISCED5B >	0000100
		6 <ISCED5A >	0000010
		Missing	0000001
Father's highest qualifications?	FISCED	0 <none >	1000000
		1 <ISCED1 >	0100000
		2 <ISCED2 >	0010000
		3 <ISCED3B,C>	0001000
		4 <ISCED3A >	0000000
		5 <ISCED5B >	0000100
		6 <ISCED5A >	0000010
		Missing	0000001
Expected educational level of student (ISCED)	SISCED	0 <ISCED1 >	100000
		1 <ISCED2 >	010000
		2 <ISCED3B,C>	001000
		3 <ISCED3A >	000000
		4 <ISCED5B >	000100
		5 <ISCED5A >	000010
		Missing	000001
Who usually lives at <home> with you? - Q4	FAMSTRUC	1 <Single parent family >	1000
		2 <Nuclear family>	0000
		3 <Mixed family>	0100
		4 <Other>	0010
		Missing	0001
Minutes of mathematics per week	MMINS	0	0 0
		1	1 0
		2	2 0
		…	…
		90000	90000 0
		Missing	(mean) 1
Total minutes of instructional time per week	TMINS	0	0 0
		1	1 0
		2	2 0
		…	…
		90000	90000 0
		Missing	(mean) 1
Mother's main job	BMMJ	1	1 0
		2	2 0
		…	…
		90	90 0
		Missing	(mean) 1
Father's main job	BFMJ	1	1 0
		2	2 0
		…	…

Conditioning variables	Variable name(s)	Variable coding	Contrast coding
		90	90 0
		Missing	(mean) 1
Student's expected job	BSMJ	1	1 0
		2	2 0
		…	…
		90	90 0
		Missing	(mean) 1
Student questionnaire WLE estimates			
Computer facilities at home (WLE)	COMPHOME	Value (decimal)	(copy) 0
		Missing	(mean) 1
Cultural possessions of the family (WLE)	CULTPOSS	Value (decimal)	(copy) 0
		Missing	(mean) 1
Home educational resources (WLE)	HEDRES	Value (decimal)	(copy) 0
		Missing	(mean) 1
Attitudes towards school (WLE)	ATSCHL	Value (decimal)	(copy) 0
		Missing	(mean) 1
Student-teacher relations at school (WLE)	STUREL	Value (decimal)	(copy) 0
		Missing	(mean) 1
Sense of belonging to school (WLE)	BELONG	Value (decimal)	(copy) 0
		Missing	(mean) 1
Interest in mathematics (WLE)	INTMAT	Value (decimal)	(copy) 0
		Missing	(mean) 1
Instrumental motivation in mathematics (WLE)	INSTMOT	Value (decimal)	(copy) 0
		Missing	(mean) 1
Mathematics self-efficacy (WLE)	MATHEFF	Value (decimal)	(copy) 0
		Missing	(mean) 1
Mathematics anxiety (WLE)	ANXMAT	Value (decimal)	(copy) 0
		Missing	(mean) 1
Mathematics self-concept (WLE)	SCMAT	Value (decimal)	(copy) 0
		Missing	(mean) 1
Control strategies (WLE)	CSTRAT	Value (decimal)	(copy) 0
		Missing	(mean) 1
Elaboration strategies (WLE)	ELAB	Value (decimal)	(copy) 0
		Missing	(mean) 1
Memorisation strategies (WLE)	MEMOR	Value (decimal)	(copy) 0
		Missing	(mean) 1
Competitive learning (WLE)	COMPLRN	Value (decimal)	(copy) 0
		Missing	(mean) 1
Co-operative learning (WLE)	COOPLRN	Value (decimal)	(copy) 0
		Missing	(mean) 1
Teacher support in mathematics lessons (WLE)	TEACHSUP	Value (decimal)	(copy) 0
		Missing	(mean) 1
Disciplinary climate in mathematics lessons (WLE)	DISCLIM	Value (decimal)	(copy) 0
		Missing	(mean) 1
ICT: Internet/entertainment use (WLE)	INTUSE	Value (decimal)	(copy) 0
		Missing	(mean) 1
ICT: Programs/software use (WLE)	PRGUSE	Value (decimal)	(copy) 0
		Missing	(mean) 1
ICT: Confidence in routine tasks (WLE)	ROUTCONF	Value (decimal)	(copy) 0
		Missing	(mean) 1
ICT: Confidence in internet tasks (WLE)	INTCONF	Value (decimal)	(copy) 0
		Missing	(mean) 1

Conditioning variables	Variable name(s)	Variable coding	Contrast coding
ICT: Confidence in high-level tasks (WLE)	HIGHCONF	Value (decimal) Missing	(copy)0 (mean)1
ICT: Attitudes towards computers (WLE)	ATTCOMP	Value (decimal) Missing	(copy)0 (mean)1

Student booklet ID	Variable name(s)	Variable coding	Contrast coding
Booklet ID	BOOKID	1	1000000000000
		2	0100000000000
		3	0010000000000
		4	0001000000000
		5	0000100000000
		6	0000010000000
		7	0000001000000
		8	0000000100000
		9	0000000010000
		10	0000000001000
		11	0000000000100
		12	0000000000010
		13	0000000000001
		14	0000000000000

Appendix 11

SCALE RELIABILITIES BY COUNTRY

OECD Member-ship	Country	Major domain: mathematics					Minor domains		
		Combined	Quantity	Space and shape	Change and relation-ships	Uncertainty	Reading	Science	Problem solving
OECD countries	Australia	0.91	0.84	0.90	0.90	0.87	0.83	0.84	0.86
	Austria	0.92	0.83	0.91	0.89	0.89	0.87	0.88	0.87
	Belgium	0.93	0.86	0.91	0.90	0.91	0.86	0.85	0.88
	Canada	0.89	0.82	0.88	0.89	0.86	0.83	0.83	0.84
	Czech Republic	0.91	0.83	0.88	0.89	0.88	0.84	0.82	0.87
	Denmark	0.90	0.84	0.89	0.90	0.85	0.82	0.82	0.85
	Finland	0.89	0.81	0.86	0.87	0.86	0.80	0.80	0.81
	France	0.90	0.86	0.89	0.89	0.88	0.84	0.84	0.85
	Germany	0.93	0.87	0.91	0.91	0.91	0.88	0.88	0.89
	Greece	0.89	0.84	0.87	0.86	0.86	0.78	0.76	0.79
	Hungary	0.90	0.84	0.87	0.90	0.88	0.81	0.81	0.86
	Iceland	0.90	0.85	0.88	0.89	0.87	0.83	0.82	0.85
	Ireland	0.91	0.83	0.90	0.90	0.88	0.87	0.85	0.87
	Italy	0.91	0.86	0.90	0.90	0.89	0.85	0.84	0.87
	Japan	0.91	0.85	0.89	0.90	0.88	0.84	0.85	0.85
	Korea	0.91	0.85	0.90	0.91	0.89	0.83	0.84	0.86
	Luxembourg	0.90	0.82	0.89	0.90	0.88	0.86	0.85	0.86
	Mexico	0.86	0.80	0.85	0.81	0.85	0.78	0.72	0.78
	Netherlands	0.93	0.88	0.92	0.92	0.90	0.87	0.88	0.89
	New Zealand	0.92	0.88	0.91	0.91	0.91	0.86	0.86	0.88
	Norway	0.90	0.85	0.89	0.87	0.87	0.82	0.81	0.85
	Poland	0.90	0.84	0.88	0.89	0.88	0.82	0.81	0.84
	Portugal	0.90	0.84	0.88	0.89	0.88	0.84	0.81	0.84
	Slovak Republic	0.91	0.84	0.89	0.89	0.88	0.83	0.82	0.86
	Spain	0.89	0.83	0.87	0.88	0.87	0.81	0.80	0.83
	Sweden	0.90	0.85	0.89	0.89	0.88	0.83	0.81	0.83
	Switzerland	0.91	0.85	0.89	0.89	0.88	0.84	0.84	0.87
	Turkey	0.91	0.86	0.89	0.89	0.89	0.82	0.83	0.87
	United Kingdom	0.92	0.85	0.91	0.91	0.88	0.86	0.87	0.87
	United States	0.92	0.85	0.91	0.92	0.89	0.87	0.85	0.88
	OECD median	0.91	0.84	0.89	0.89	0.88	0.83	0.83	0.86
Partner countries	Brazil	0.88	0.84	0.88	0.85	0.88	0.76	0.73	0.81
	Hong Kong-China	0.92	0.88	0.91	0.91	0.88	0.85	0.85	0.88
	Indonesia	0.83	0.76	0.82	0.81	0.83	0.70	0.68	0.71
	Latvia	0.89	0.85	0.88	0.90	0.87	0.80	0.78	0.82
	Liechtenstein	0.91	0.85	0.89	0.91	0.88	0.84	0.84	0.87
	Macao-China	0.88	0.83	0.89	0.87	0.87	0.78	0.79	0.83
	Russian Federation	0.88	0.84	0.87	0.88	0.86	0.76	0.74	0.81
	Serbia	0.88	0.80	0.86	0.85	0.86	0.78	0.76	0.79
	Thailand	0.86	0.81	0.84	0.85	0.85	0.76	0.74	0.77
	Tunisia	0.85	0.81	0.84	0.81	0.83	0.72	0.69	0.72
	Uruguay	0.89	0.84	0.86	0.86	0.87	0.77	0.75	0.78
	Partner median	0.88	0.84	0.87	0.86	0.87	0.77	0.75	0.81

Appendix 12

DETAILS OF THE MATHEMATICS ITEMS USED IN PISA 2003

Identification Number	Name	Source	Language in which submitted	Scale	Cluster	International % correct	Item parameters (RP=0.5)				Thresholds (RP=0.62 PISA scale)		
							Difficulty	Tau-1	Tau-2	Tau-3	1	2	3
M033Q01	P2000 A View Room	CITO	Dutch	Space and shape	M1	76.77 (0.22)	-1.496				432		
M034Q01T	P2000 Bricks	CITO	Dutch	Space and shape	M2	43.27 (0.27)	0.432				582		
M124Q01	P2000 Walking	CITO	Dutch	Change and relationships	M3	36.34 (0.32)	0.797				611		
M124Q03T	P2000 Walking	CITO	Dutch	Change and relationships	M3	20.62 (0.21)	1.488	-0.301	0.076	0.225	605	666	723
M144Q01T	P2000 Cube Painting	USA	English	Space and shape	M4	62.09 (0.25)	-0.666				497		
M144Q02T	P2000 Cube Painting	USA	English	Space and shape	M4	27.44 (0.22)	1.235				645		
M144Q03	P2000 Cube Painting	USA	English	Space and shape	M4	75.16 (0.22)	-1.491				432		
M144Q04T	P2000 Cube Painting	USA	English	Space and shape	M4	38.42 (0.26)	0.641				599		
M145Q01T	P2000 Cubes	CITO	Dutch	Space and shape	M2	68.03 (0.27)	-0.906				478		
M150Q01	P2000 Growing Up	CITO	Dutch	Change and relationships	M5	66.96 (0.25)	-0.913				477		
M150Q02T	P2000 Growing Up	CITO	Dutch	Change and relationships	M5	68.77 (0.24)	-0.979	-0.384	0.384		420	525	
M150Q03T	P2000 Growing Up	CITO	Dutch	Change and relationships	M5	44.83 (0.25)	0.322				574		
M155Q01	P2000 Population Pyramids	CITO	Dutch	Change and relationships	M6	64.86 (0.30)	-0.891				479		
M155Q02T	P2000 Population Pyramids	CITO	Dutch	Change and relationships	M6	60.66 (0.25)	-0.480	0.682	-0.682		492	531	
M155Q03T	P2000 Population Pyramids	CITO	Dutch	Change and relationships	M6	16.79 (0.17)	1.616	0.197	-0.197		643	706	
M155Q04T	P2000 Population Pyramids	CITO	Dutch	Change and relationships	M6	56.49 (0.27)	-0.391				518		
M179Q01T	P2000 Robberies	TIMSS	English	Uncertainty	M1	29.50 (0.22)	1.114	-0.503	0.503		577	694	
M192Q01T	P2000 Containers	Germany	German	Change and relationships	M2	40.41 (0.29)	0.578				594		
M266Q01T	P2000 Carpenter	Australia	English	Space and shape	M7	19.95 (0.20)	1.782				687		
M273Q01T	P2000 Pipelines	Czech Republic	Czech	Space and shape	M7	54.92 (0.28)	-0.307				525		
M302Q01T	Car Drive	TIMSS	English	Change and relationships	M7	95.32 (0.13)	-3.680				262		
M302Q02	Car Drive	TIMSS	English	Change and relationships	M7	78.42 (0.23)	-1.725				414		
M302Q03	Car Drive	TIMSS	English	Change and relationships	M7	30.00 (0.26)	1.055				631		
M305Q01	Map	ACER	English	Space and shape	M3	64.14 (0.25)	-0.648				498		
M402Q01	Internet Relay Chat	ACER	English	Change and relationships	M1	53.72 (0.26)	-0.204				533		
M402Q02	Internet Relay Chat	ACER	English	Change and relationships	M1	28.79 (0.24)	1.119				636		
M406Q01	Running Tracks	ACER	English	Space and shape	M5	28.66 (0.25)	1.163				639		
M406Q02	Running Tracks	ACER	English	Space and shape	M5	19.33 (0.23)	1.775				687		
M406Q03	Running Tracks	ACER	English	Space and shape	M5	18.72 (0.21)	1.845				692		
M408Q01T	Lotteries	ACER	English	Uncertainty	M2	41.60 (0.24)	0.494				587		
M411Q01	Diving	ACER	English	Quantity	M5	51.39 (0.30)	-0.046				545		
M411Q02	Diving	ACER	English	Uncertainty	M5	45.99 (0.29)	0.201				564		
M413Q01	Exchange Rate	ACER	English	Quantity	M5	79.66 (0.25)	-1.833				406		
M413Q02	Exchange Rate	ACER	English	Quantity	M5	73.86 (0.28)	-1.408				439		
M413Q03T	Exchange Rate	ACER	English	Quantity	M5	40.34 (0.27)	0.474				586		
M420Q01T	Transport	ACER	English	Uncertainty	M6	49.87 (0.27)	-0.059				544		
M421Q01	Height	ACER	English	Uncertainty	M4	64.97 (0.32)	-0.812				485		
M421Q02T	Height	ACER	English	Uncertainty	M4	17.85 (0.21)	1.982				703		
M421Q03	Height	ACER	English	Uncertainty	M4	38.04 (0.27)	0.680				602		
M423Q01	Tossing Coins	ACER	English	Uncertainty	M2	81.66 (0.19)	-1.821				407		
M434Q01	Room Numbers	ACER	English	Quantity	M3								
M438Q01	Exports	Argentina	Spanish	Uncertainty	M3	78.69 (0.20)	-1.567				427		
M438Q02	Exports	Argentina	Spanish	Uncertainty	M3	48.33 (0.32)	0.213				565		

Identification Number	Name	Source	Language in which submitted	Scale	Cluster	International % correct	Item parameters (RP=0.5)				Thresholds (RP=0.62 PISA scale)		
							Difficulty	Tau-1	Tau-2	Tau-3	1	2	3
M442Q02	Braille	ACER	English	Quantity	M6	41.78 (0.26)	0.376				578		
M446Q01	Thermometer Cricket	ACER	English	Change and relationships	M2	68.22 (0.28)	-1.012				470		
M446Q02	Thermometer Cricket	ACER	English	Change and relationships	M2	6.79 (0.14)	3.239				801		
M447Q01	Tile Arrangement	ACER	English	Space and shape	M6	70.23 (0.28)	-1.120				461		
M462Q01T	Third Side	Sweden	English	Space and shape	M3	14.11 (0.18)	1.971	0.185	-0.185		671	734	
M464Q01T	The Fence	Sweden	English	Space and shape	M1	25.11 (0.25)	1.462				662		
M467Q01	Coloured Candies	Canada	English	Uncertainty	M1	50.21 (0.27)	0.001				549		
M468Q01T	Science Tests	Canada	English	Uncertainty	M6	46.77 (0.32)	0.101				556		
M474Q01	Running Time	Canada	English	Quantity	M3	74.07 (0.22)	-1.246				452		
M484Q01T	Bookshelves	Czech Republic	English	Quantity	M6	60.88 (0.26)	-0.639				499		
M496Q01T	Cash Withdrawal	ACER	English	Quantity	M6	53.12 (0.29)	-0.196				533		
M496Q02	Cash Withdrawal	ACER	English	Quantity	M6	65.65 (0.23)	-0.896				479		
M505Q01	Litter	CITO	English	Uncertainty	M3	51.55 (0.26)	0.027				551		
M509Q01	Earthquake	CITO	English	Uncertainty	M6	46.48 (0.22)	0.110				557		
M510Q01T	Choices	CITO	English	Quantity	M3	48.76 (0.29)	0.139				559		
M513Q01	Test Scores	CITO	English	Uncertainty	M7	32.21 (0.25)	0.913				620		
M520Q01T	Skateboard	CITO	English	Quantity	M2	72.01 (0.25)	-0.879	0.865	-0.865		464	496	
M520Q02	Skateboard	CITO	English	Quantity	M2	45.53 (0.26)	0.272				570		
M520Q03T	Skateboard	CITO	English	Quantity	M2	49.78 (0.29)	0.074				554		
M547Q01T	Staircase	Norway	English	Space and shape	M3	78.04 (0.20)	-1.640				421		
M555Q02T	Number Cubes	Norway	English	Space and shape	M2	62.97 (0.26)	-0.582				503		
M559Q01	Telephone Rates	Italy	English	Quantity	M4	61.00 (0.28)	-0.569				504		
M564Q01	Chair Lift	Italy	English	Quantity	M1	49.26 (0.29)	0.026				551		
M564Q02	Chair Lift	Italy	English	Uncertainty	M1	45.56 (0.27)	0.257				569		
M571Q01	Stop The Car	Germany	German	Change and relationships	M4	48.83 (0.28)	0.098				556		
M598Q01	Making A Booklet	Switzerland	German	Space and shape	M5	64.15 (0.23)	-0.773				488		
M603Q01T	Number Check	Austria	German	Quantity	M7	47.10 (0.23)	0.130				559		
M603Q02T	Number Check	Austria	German	Quantity	M7	36.08 (0.29)	0.668				601		
M702Q01	Support For President	NIER	Japanese	Uncertainty	M2	35.66 (0.30)	0.846				615		
M704Q01T	The Best Car	NIER	Japanese	Change and relationships	M4	72.91 (0.24)	-1.307				447		
M704Q02T	The Best Car	NIER	Japanese	Change and relationships	M4	25.42 (0.21)	1.389				657		
M710Q01	Forecast of Rain	NIER	Japanese	Uncertainty	M5	33.88 (0.26)	0.919				620		
M800Q01	Computer Game	Canada	English	Quantity	UHM	91.77 (0.12)	-3.077				309		
M803Q01T	Labels	Canada	English	Uncertainty	M7	28.14 (0.23)	1.188				641		
M806Q01T	Step Pattern	Canada	French	Quantity	M3	66.19 (0.23)	-0.824				484		
M810Q01T	Bicycles	Canada	English	Quantity	M1	68.31 (0.22)	-0.968				473		
M810Q02T	Bicycles	Canada	English	Quantity	M1	71.71 (0.24)	-1.156				459		
M810Q03T	Bicycles	Canada	English	Change and relationships	M1	20.14 (0.20)	1.563	-0.052	0.052		631	631	
M828Q01	Carbon Dioxide	Netherlands	English	Change and relationships	M7	39.74 (0.25)	0.573				593		
M828Q02	Carbon Dioxide	Netherlands	English	Uncertainty	M7	54.26 (0.25)	-0.196				533		
M828Q03	Carbon Dioxide	Netherlands	English	Quantity	M7	32.08 (0.25)	1.033				629		
M833Q01T	Seeing the Tower	Netherlands	English	Space and shape	M1	31.81 (0.24)	1.023				628		

Appendix 13

DETAILS OF THE READING ITEMS USED IN PISA 2003

Identi-ficaion Number	Name	Source	Language in which submitted	Scale	Cluster	Interna-tional % correct	Item parameters (RP=0.5)				Thresholds (RP=0.62 PISA scale)		
							Diffi-culty	Tau-1	Tau-2	Tau-3	1	2	3
R055Q01	Drugged Spiders	Cito	English	Interpreting	R2	81.38 (0.19)	-1.27				401		
R055Q02	Drugged Spiders	Cito	English	Reflecting	R2	47.73 (0.26)	0.63				554		
R055Q03	Drugged Spiders	Cito	English	Interpreting	R2	58.83 (0.35)	0.27				525		
R055Q05	Drugged Spiders	Cito	English	Interpreting	R2	72.44 (0.31)	-0.69				448		
R067Q01	Aesop	Greece	Greek	Interpreting	R1	89.20 (0.18)	-2.08				336		
R067Q04	Aesop	Greece	Greek	Reflecting	R1	56.38 (0.23)	0.25	-0.437	0.437		466	581	
R067Q05	Aesop	Greece	Greek	Reflecting	R1	66.47 (0.27)	-0.18	0.578	-0.578		466	511	
R102Q04A	Shirts	Cito	English	Interpreting	R1	31.31 (0.27)	1.53				626		
R102Q05	Shirts	Cito	English	Interpreting	R1	43.71 (0.30)	0.87				573		
R102Q07	Shirts	Cito	English	Interpreting	R1	81.96 (0.23)	-1.42				389		
R104Q01	Telephone	New Zealand	English	Retrieving information	R2	82.95 (0.25)	-1.47				385		
R104Q02	Telephone	New Zealand	English	Retrieving information	R2	34.18 (0.27)	1.44				619		
R104Q05	Telephone	New Zealand	English	Retrieving information	R2	24.84 (0.20)	2.17	-1.111	1.111		581	774	
R111Q01	Exchange	Finland	Finnish	Interpreting	R2	64.52 (0.26)	-0.19				488		
R111Q02B	Exchange	Finland	Finnish	Reflecting	R2	33.25 (0.19)	1.54	-0.685	0.685		556	697	
R111Q06B	Exchange	Finland	Finnish	Reflecting	R2	42.96 (0.26)	0.89	0.782	-0.782		557	593	
R219Q01T	Employment	IALS	IALS	Interpreting	R1	69.38 (0.26)	-0.59				456		
R219Q01E	Employment	IALS	IALS	Retrieving information	R1	57.31 (0.31)	0.10				511		
R219Q02	Employment	IALS	IALS	Reflecting	R1	78.02 (0.23)	-1.13				413		
R220Q01	South Pole	France	French	Retrieving information	R1	42.77 (0.31)	0.86				572		
R220Q02B	South Pole	France	French	Interpreting	R1	62.95 (0.27)	-0.14				492		
R220Q04	South Pole	France	French	Interpreting	R1	61.44 (0.29)	-0.10				495		
R220Q05	South Pole	France	French	Interpreting	R1	82.57 (0.24)	-1.38				392		
R220Q06	South Pole	France	French	Interpreting	R1	66.18 (0.28)	-0.34				476		
R227Q01	Optician	Switzerland	German	Interpreting	R2	53.58 (0.29)	0.40				535		
R227Q02T	Optician	Switzerland	German	Retrieving information	R2	57.07 (0.24)	0.16	-1.076	1.076		422	611	
R227Q03	Optician	Switzerland	German	Reflecting	R2	53.77 (0.30)	0.46				540		
R227Q06	Optician	Switzerland	German	Retrieving information	R2	70.95 (0.31)	-0.56				459		

Appendix 14

DETAILS OF THE SCIENCE ITEMS USED IN PISA 2003

Identification Number	Name	Source	Language in which submitted	Cluster	International % correct	Item parameters (RP=0.5)				Thresholds (RP=0.62 PISA scale)		
						Difficulty	Tau-1	Tau-2	Tau-3	1	2	3
S114Q03T	P2000 Greenhouse	CITO	Dutch	S1	54.02 (0.28)	-0.290				527.7		
S114Q04T	P2000 Greenhouse	CITO	Dutch	S1	35.99 (0.22)	0.544	-0.078	0.078		556.1	650.7	
S114Q05T	P2000 Greenhouse	CITO	Dutch	S1	22.26 (0.24)	1.480				688.4		
S128Q01	P2000 Cloning	CITO	French	S2	64.67 (0.23)	-0.661				494.0		
S128Q02	P2000 Cloning	CITO	French	S2	48.68 (0.24)	0.202				572.3		
S128Q03T	P2000 Cloning	CITO	French	S2	62.09 (0.29)	-0.523				506.5		
S129Q01	P2000 Daylight	ACER	English	S2	42.60 (0.23)	0.423				592.4		
S129Q02T	P2000 Daylight	ACER	English	S2	18.61 (0.18)	1.535	0.519	-0.519		666.7	720.2	
S131Q02T	P2000 Good Vibrations	ACER	English	S2	46.41 (0.28)	0.263				577.9		
S131Q04T	P2000 Good Vibrations	ACER	English	S2	26.11 (0.22)	1.409				681.9		
S133Q01	P2000 Research	USA	English	S1	60.72 (0.22)	-0.596				499.9		
S133Q03	P2000 Research	USA	English	S1	36.60 (0.25)	0.642				612.2		
S133Q04T	P2000 Research	USA	English	S1	45.43 (0.28)	0.133				566.1		
S213Q01T	P2000 Clothes	Australia	English	S1	41.98 (0.25)	0.359				586.6		
S213Q02	P2000 Clothes	Australia	English	S1	76.22 (0.24)	-1.455				421.9		
S252Q01	P2000 South Rainea	Korea	Korean	S1	52.07 (0.25)	-0.181				537.6		
S252Q02	P2000 South Rainea	Korea	Korean	S1	68.71 (0.22)	-0.965				466.3		
S252Q03T	P2000 South Rainea	Korea	Korean	S1	58.06 (0.22)	-0.465				511.8		
S256Q01	P2000 Spoons	TIMSS	English	S2	87.12 (0.18)	-2.212				353.2		
S268Q01	P2000 Algae	Australia	English	S2	71.30 (0.23)	-1.099				454.2		
S268Q02T	P2000 Algae	Australia	English	S2	37.08 (0.31)	0.800				626.6		
S268Q06	P2000 Algae	Australia	English	S2	55.86 (0.28)	-0.168				538.7		
S269Q01	P2000 Earth's Temperature	CITO	Dutch	S2	59.65 (0.30)	-0.456				512.6		
S269Q03T	P2000 Earth's Temperature	CITO	Dutch	S2	40.42 (0.25)	0.557				604.5		
S269Q04T	P2000 Earth's Temperature	CITO	Dutch	S2	35.76 (0.20)	0.887				634.4		
S304Q01	Water	CITO	Dutch	S2	44.94 (0.31)	0.327				583.6		
S304Q02	Water	CITO	Dutch	S2	61.83 (0.26)	-0.573				501.9		
S304Q03a	Water	CITO	Dutch	S2	38.02 (0.26)	0.708				618.3		
S304Q03b	Water	CITO	Dutch	S2	50.15 (0.29)	0.069				560.2		
S326Q01	Milk	CITO	Dutch	S1	58.33 (0.29)	-0.412				516.5		
S326Q02	Milk	CITO	Dutch	S1	62.65 (0.27)	-0.682				492.1		
S326Q03	Milk	CITO	Dutch	S1	56.73 (0.28)	-0.468				511.5		
S326Q04T	Milk	CITO	Dutch	S1	22.33 (0.23)	1.480				688.3		
S327Q01T	Tidal Energy	CITO	Dutch	S1	61.52 (0.28)	-0.611				498.5		
S327Q02	Tidal Energy	CITO	Dutch	S1								

© OECD 2005

Appendix 15

DETAILS OF THE PROBLEM-SOLVING ITEMS USED IN PISA 2003

Identification Number	Name	Source	Language in which submitted	Cluster	International % correct	Item parameters (RP=0.5)				Thresholds (RP=0.62 PISA scale)		
						Difficulty	Tau-1	Tau-2	Tau-3	1	2	3
X402Q01T	Library System	ACER	English	PS2	74.80 (0.25)	-1.33				436.9		
X402Q02T	Library System	ACER	English	PS2	14.32 (0.16)	1.48	1.09	0.15	-1.23	658.0	676.9	693.2
X412Q01	Design by Numbers	ACER	English	PS1	50.29 (0.25)	-0.07				544.3		
X412Q02	Design by Numbers	ACER	English	PS1	48.28 (0.27)	0.03				552.7		
X412Q03	Design by Numbers	ACER	English	PS1	39.58 (0.25)	0.42	1.04	-1.04		570.5	600.5	
X414Q01	Course Design	ACER	English	PS2	31.06 (0.23)	0.77	1.14	-1.14		602.4	629.4	
X415Q01T	Transit System	ACER	English	PS2	24.15 (0.17)	1.37	-0.40	0.40		608.1	725.5	
X417Q01	Children's Camp	Leeds	English	PS1	40.09 (0.23)	0.47	-0.44	0.44		529.0	650.4	
X423Q01T	Freezer	CITO	English	PS1	49.21 (0.26)	0.02				551.3		
X423Q02T	Freezer	CITO	English	PS1	44.62 (0.27)	0.28				573.4		
X430Q01	Energy Needs	Leeds	English	UHPS	84.79 (0.21)	-2.23				360.6		
X430Q02	Energy Needs	Leeds	English	UHPS	32.07 (0.26)	0.65	0.83	-0.83		586.9	623.7	
X601Q01T	Cinema Outing	Leeds	English	PS1	67.21 (0.25)	-0.80	0.02	-0.02		441.6	522.0	
X601Q02	Cinema Outing	Leeds	English	PS1	68.06 (0.26)	-0.96				468.2		
X602Q01	Holiday	Leeds	English	PS2	45.87 (0.24)	0.24				570.1		
X602Q02	Holiday	Leeds	English	PS2	35.63 (0.27)	0.57	2.09	-2.09		592.8	603.5	
X603Q01	Irrigation	Leeds	English	PS2	62.88 (0.25)	-0.62				497.3		
X603Q02T	Irrigation	Leeds	English	PS2	51.34 (0.28)	-0.08				543.5		
X603Q03	Irrigation	Leeds	English	PS2	54.44 (0.28)	-0.22				531.6		

Appendix 16

LEVELS OF PARENTAL EDUCATION CONVERTED INTO YEARS OF SCHOOLING

		Did not go to school	Completed <ISCED Level 1 (primary education)>	Completed <ISCED Level 2 (lower secondary education)>	Completed <ISCED Levels 3B or 3C (upper secondary education aimed at direct entry into the labour market)>	Completed <ISCED Level 3A (upper secondary education aimed at entry into tertiary education)>	Completed <ISCED Level 5A (tertiary education)>	Completed <ISCED Level 5B (tertiary educa-tion)>
OECD countries	Australia	0.0	6.5	10.0	11.5	12.0	15.0	14.0
	Austria	0.0	4.0	8.0	11.0	13.0	17.0	15.0
	Belgium	0.0	6.0	8.0	12.0	12.0	16.0	15.0
	Canada	0.0	6.0	9.0	12.0	12.0	17.0	15.0
	Czech Republic	0.0	5.0	9.0	12.0	13.0	17.0	16.0
	Denmark	0.0	6.0	9.5	12.5	12.5	16.5	15.5
	Finland	0.0	6.0	9.0	12.0	12.0	15.5	14.5
	France	0.0	5.0	9.0	11.0	12.0	15.0	14.0
	Germany	0.0	4.0	10.0	12.0	12.5	17.0	15.0
	Greece	0.0	6.0	9.0	11.5	12.0	17.0	15.5
	Hungary	0.0	4.0	8.0	10.5	12.0	16.5	13.5
	Iceland	0.0	7.0	10.0	13.0	14.0	17.0	16.5
	Ireland	0.0	6.0	9.0	a	12.0	16.0	14.0
	Italy	0.0	5.0	8.0	11.0	13.0	17.0	16.0
	Japan	0.0	6.0	9.0	12.0	12.0	16.0	14.0
	Korea	0.0	6.0	9.0	12.0	12.0	16.0	15.0
	Luxembourg	0.0	6.0	9.0	12.0	13.0	17.0	17.0
	Mexico	0.0	6.0	9.0	12.0	12.0	16.0	14.0
	Netherlands	0.0	6.0	10.0	a	12.0	15.0	a
	New Zealand	0.0	6.0	10.0	12.0	13.0	16.0	16.0
	Norway	0.0	7.0	10.0	13.0	13.0	17.0	15.0
	Poland	0.0	a	8.0	11.0	12.0	16.0	15.0
	Portugal	0.0	6.0	9.0	12.0	12.0	17.0	15.0
	Slovak Republic	0.0	4.0	9.0	12.0	12.5	17.0	15.0
	Spain	0.0	6.0	10.0	12.0	12.0	15.0	14.0
	Sweden	0.0	6.0	9.0	12.0	12.0	15.5	14.0
	Switzerland	0.0	6.0	9.0	12.0	12.5	15.0	14.0
	Turkey	0.0	5.0	8.0	11.0	11.0	16.0	14.0
	United Kindgom	0.0	6.0	9.0	11.0	12.0	16.0	15.0
	United States	0.0	6.0	9.0	a	12.0	16.0	15.0
Partner countries	Brazil	0.0	4.0	8.0	11.0	11.0	16.0	14.5
	Hong Kong-China	0.0	6.0	9.0	11.0	13.0	16.0	14.0
	Indonesia	0.0	6.0	9.0	12.0	12.0	16.0	15.0
	Latvia	0.0	4.0	9.0	12.0	12.0	16.0	16.0
	Liechtenstein	0.0	5.0	9.0	11.0	12.0	15.0	14.0
	Macao-China	0.0	6.0	9.0	11.0	13.0	16.0	14.0
	Russian Federation	0.0	4.0	9.0	11.5	12.0	15.0	a
	Serbia	0.0	4.0	8.0	11.0	12.0	16.0	14.0
	Thailand	0.0	6.0	9.0	12.0	12.0	16.0	14.0
	Tunisia	0.0	6.0	9.0	12.0	13.0	17.0	16.0
	Uruguay	0.0	6.0	9.0	11.0	12.0	16.0	15.0

a. The category does not apply in the country concerned. Data are therefore missing.

Appendix 17

STUDENT LISTING FORM

Page ___ of ___

School ID: _____ Country Name: _____

School Name: _____ List Prepared By: _____

Address: _____ Telephone #: _____

_____ Date List Prepared: _____

_____ Total # Students Listed: _____

DIRECTIONS: PLEASE COMPLETE COLUMNS A, B, C, AND D, FOR EVERY STUDENT <BORN IN 1984>.

Include students who may be excluded from other testing programs, such as some students with disabilities or limited language proficiency. Detailed instructions and information about providing computer-generated lists are on the other side of this page.

For Sampling Only		(A)	(B)	(C)	(D)
Selected Student (Enter "S")	Line #	Student's Name (First Middle Initial Last)	Grade	Sex (M / F)	Birth Date (mm/yy)

A. Instructions for Preparing a List of Eligible Students

1. Please prepare a list of **ALL students <born in 1984. . .NPM must insert eligibility criteria>** using the most current enrollment records available.
2. Include on the list students who typically may be excluded from other testing programs (such as some students with disabilities or limited language proficiency).
3. Write the name for each eligible student. Please also specify current grade, sex, and birth date for each student.
4. If confidentiality is a concern in listing student names, then a unique student identifier may be substituted. Because some students may have the same or similar names, it is important to include a birth date for each student.
5. The list may be computer-generated or prepared manually using the PISA Student Listing Form. A Student Listing Form is on the reverse side of these instructions. You may copy this form or request copies from your National Project Manager.
6. If you use the Student Listing Form on the reverse side of this page, do **not** write in the "For Sampling Only" columns.
7. Send the list to the National Project Manager (NPM) to arrive no later than **<insert DATE>**. Please address to the NPM as follows: **<insert name and mailing address>**

C. Suggestions for Preparing Computer-generated Lists

- Write the school name and address on list.
- List students in alphabetical order.
- Number the students.
- Double-space the list.
- Allow left-hand margin of at least two inches.
- Include the date the printout was prepared.
- Define any special codes used.
- Include preparer's name and telephone number.

Notes

1 Note that the original item asked several questions about this stimulus that have not been included here.

2 Please retain original documentation as evidence of authorisation.