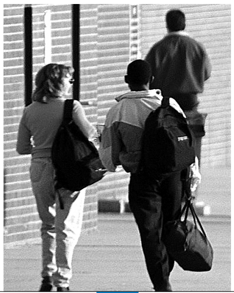


15

Proficiency Scale Construction

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

INTRODUCTION

The PISA test design makes it possible to use techniques of modern item response modelling (see Chapter 9) 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 overall PISA literacy scale in the relevant test domain of reading, mathematics or science.

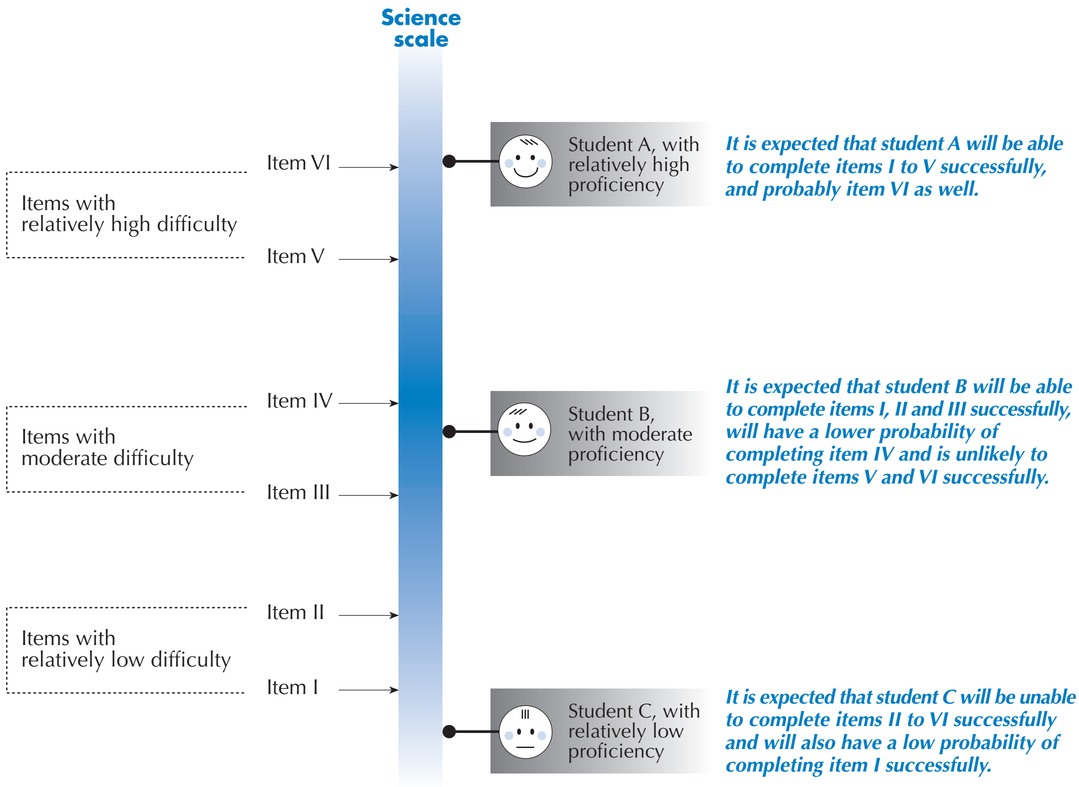
PISA assesses students, and uses the outcomes of that assessment to produce estimates of students' proficiency in relation to a number of literacy variables. These variables are defined in the relevant PISA literacy framework (OECD, 2006). For each of these literacy variables, one or more scales are defined, which stretch from very low levels of literacy through to very high levels. When thinking about what such a scale means in terms of 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 15.1 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 and was revised in the lead-up to the PISA 2003 survey. Essentially the same methodology has again been used to develop proficiency descriptions for PISA 2006. Given the volume and breadth of data that were available from the PISA 2006 assessment, development of more detailed descriptions of scientific literacy became possible. 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 and 2006. The detailed descriptions used for mathematics in 2003 were used again in 2006.

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Figure 15.1
The relationship between items and students on a proficiency scale



The SEG worked with the consortium to develop sets of described proficiency scales for PISA science. Consultations regarding these described scales with the PGB, the science forum, NPMs and the PISA TAG 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 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 scales

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, given the small number of items in PISA 2000 and 2003, a single overall proficiency scale was developed to report results. However, as with mathematics in 2003, the expanded focus on science in 2006 allowed for a division into scales for reporting purposes. Two forms of scale were considered. One of these was based in definitions of scientific competencies involving the identification of scientific issues, the explanation of phenomena scientifically and the use of scientific evidence. The other form separated scientific knowledge into 'knowledge of science' involving the application of scientific concepts in the major fields of phyics, chemistry, biology, Earth and space science, and technology; and 'knowledge about science' involving the central processes underpinning in the way scientists go about obtaining and using data – in other words, understanding scientific methodology. The scales finally selected for inclusion in the PISA 2006 primary database were the three competency based scales: *identifying scientific issues*, *explaining phenomena scientifically* and *using scientific evidence*.

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 also 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.

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. Science 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 sub-scale 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. Further detail on this analysis is provided below.

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.

Many items had a single achievement threshold (associated with students providing a correct rather than incorrect response). 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, and so on).

Within each scale, 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) were 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. 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 TAG, and the approach to defining levels and associating students with those levels that had been used in the reporting of PISA 2000 and PISA 2003 results was applied.

Stage 7: Validating

Two major approaches to validation were then considered by the science 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. 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 we divide the proficiency continuum up into levels that might have some utility? And having defined levels, how should we decide on the level to which a particular student should be assigned? 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 we insist 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 would we expect the student 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 we expect 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 and 2003 was essentially retained for PISA 2006.

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 very practical matter of interpreting what it means to be at a level, and has very 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 they enable us to communicate 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 mean that we can have 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% 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% 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% 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 used again in PISA 2003 and 2006. This method links the two variables mentioned in the preceding paragraphs, 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% 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").

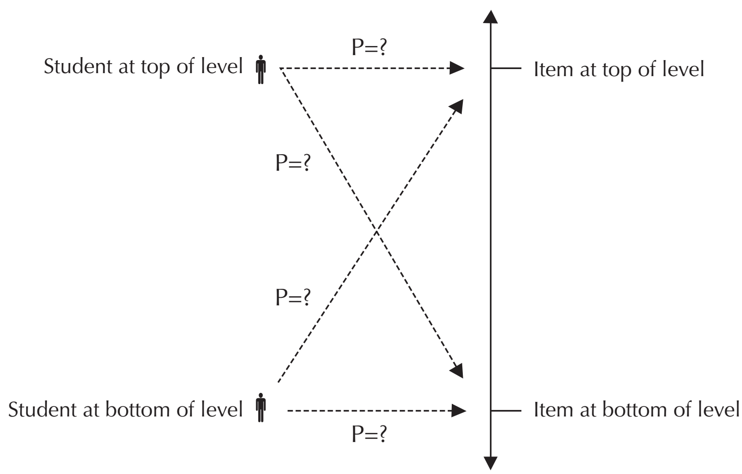
Figure 15.2 summarises the relationship among these three mathematically linked variables. It shows a vertical line representing a part of the scale being defined, one of the bounded levels on the scale, a student at both the top and the bottom of the level, and reference to an item at the top and an 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% of the items in a test at that 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% at the bottom of the level except for the lowest described level.

The highest and lowest 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.

Figure 15.2
What it means to be at a level



As levels 2, 3, 4 and 5 (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 SCIENCE

In this section, the way in which levels of scientific literacy are defined, described and reported will be discussed. Levels of performance on the PISA scientific literacy scale will be established and described, and they will be exemplified using a number of items from the PISA 2006 assessment.

Building an item map

The data from the PISA science assessment were processed to generate a set of item difficulty measures for the 103 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 109 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 science framework that were relevant to each item. For example, each item was analysed to determine which competency and type of knowledge (of or about) was involved in a correct response. Similarly, the situation (context) in which the stimulus and question were located, and to which the competencies, knowledge and attitudes were related, was identified. This included identifying whether the science involved was of personal, social or global interest. 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.

Following data analysis and the resultant generation of difficulty estimates for each of the 109 items steps, the items and item steps were associated with their difficulty estimates, with their framework classifications, and with their brief qualitative descriptions. Figure 15.3 shows a map of some of this information from a sample of items from the PISA 2006 test. Each row in Figure 15.3 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.

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 science framework's discussion of scientific literacy. Patterns emerge that make it possible to describe aspects of scientific literacy that are consistently associated with various locations along the difficulty continuum shown by the map. For example, at a very general level it can be seen that the easiest items are predominately in the *explaining phenomena scientifically* competence and lie in the *knowledge of science* area. These items are similar in that they require little interpretation, the recall of relatively straight forward factual knowledge, and the application of that knowledge in simple familiar situations. This pattern in not repeated above the mid-point of level two (defined in Table 15.1) in the sense that a specific competence is dominant. Above this level the distribution of competencies and knowledge areas is more even. This observation applies equally well to the full set of science items. However, it is possible to see growth in a number of dimensions as student student ability increases on the scientific literacy scale.

Figure 15.3
A map for selected science items

Code	Item name	Item difficulty on PISA scale	Item demands	Competency			Knowledge						Focus		
							of				about				
				Identifying scientific issues	*Explaining phenomena scientifically*	*Using scientific evidence*	*Physical systems*	*Living systems*	*Earth and space systems*	*Technology systems*	*Scientific enquiry*	*Scientific explanations*	Personal	Social	Global
S485Q05(2)	**ACID RAIN**	717	The reason for a control in an investigation is understood and explicitly recognised. An ability to understand the modelling in the investigation is a pre-requisite.	•							•		•		
S114Q05	**GREENHOUSE**	709	There is a pre-requisite to understand the need to control variables. Knowledge of factors contributing to the greenhouse effect is then applied in determining a variable to be controlled.		•				•						•
S114Q04(2)	**GREENHOUSE**	659	Given a conclusion can compare two graphs and locate corresponding areas that are at odds with that conclusion and accurately describe that difference.			•						•			•
S447Q05	**SUNSCREENS**	616	Correctly interprets a dataset expressed diagrammatically and provides an explanation that summarises the data.			•						•	•		
S447Q02	**SUNSCREENS**	588	The control 'aspects' of an investigation are recognised.	•							•		•		
S493Q05	**PHYSICAL EXERCISE**	583	Recognition that increased exercise results in increased respiration and thus the need for more oxygen and/or removal of more carbon dioxide.		•			•					•		
S114Q04(1)	**GREENHOUSE**	568	Recognises differences in two graphs relating to a phenomenon but cannot provide a clear explanation as to why the differences are at odds with a given conclusion.			•						•			•
S213Q01	**CLOTHES**	567	Can apply knowledge of the features of a scientific investigation to decisions about whether specific issues are scientifically investigatable.	•							•			•	
S493Q01	**PHYSICAL EXERCISE**	545	Can identify some features of physical exercise that are advantageous to health – cardiovascular system, bodyweight.		•			•					•		
S114Q03	**GREENHOUSE**	529	Shows an understanding of what two graphs relating to a phenomenon are depicting and can compare them for similarities.			•						•			•
S485Q05(1)	**ACID RAIN**	513	Recognises that a comparison is being made between two tests but is unable to articulate the purpose of the control.	•							•		•		
S477Q04	**MARY MONTAGU**	507	Recognises that the immune systems of young and old people are less resistant to viruses than those of the general population.		•			•						•	
S447Q03	**SUNSCREENS**	499	Can recognise the change and measured variables from a description of an investigation and as a consequence identify the question motivating the investigation.	•							•		•		
S426Q07	**GRAND CANYON**	485	Can recognise issues in which scientific measurement can be applied to answwering a question.	•							•			•	
S485Q03	**ACID RAIN**	460	Recognises that the loss of gas in a chemical reaction results in a reduction of mass for the products left behind.			•	•						•		
S426Q03	**GRAND CANYON**	451	Applies knowledge that water increases in volume as it changes from liquid to solid.		•				•					•	
S477Q03	**MARY MONTAGU**	431	Recalls knowledge of the role of antibodies in immunity.		•			•						•	
S508Q03	**GENETICALLY MODIFIED CROPS**	421	Understands that a fair test involves finding out if an outcome is affected by a range of extraneous conditions.	•							•			•	
S213Q02	**CLOTHES**	399	Can select the correct apparatus to measure an electric current.		•					•			•		
S493Q03	**PHYSICAL EXERCISE**	386	Rejects the notion that fats are formed in the muscles and knows that the rate of flow of blood increases during exercise.		•			•					•		

Based on the patterns observed when the full question set is reviewed against the three proficiency scales, it is possible to characterise the increase in the levels of complexity of competencies measured. This can be done by referring to the ways in which science competencies are associated with questions located at different points ranging from the bottom to the top of the scale. The ascending difficulty of science questions in PISA 2006 is associated with the following characteristics, which require all three competencies but which shift in emphasis as students progress from the identification of issues to the use of evidence to communicate an answer, decision or solution.

The degree to which the transfer and application of knowledge is required: At the lowest levels the application of knowledge is simple and direct. The requirement can often be fulfilled with simple recall of single facts. At higher levels of the scale, individuals are required to identify multiple fundamental concepts and combine categories of knowledge in order to respond correctly.

The degree of cognitive demand required to analyse the presented situation and synthesise an appropriate answer: This centres on features such as the depth of scientific understanding required, the range of scientific understandings required and the proximity of the situation to the students' life. At the highest level this is characterised by in-depth understanding, an ability to apply a range of scientific understandings and to apply these in broad or global contexts.

The degree of analysis needed to answer the question: This includes the demands arising from the requirement to discriminate among issues presented in the situation under analysis, identify the appropriate knowledge domain *(Knowledge of science* and *Knowledge about science)*, and use appropriate evidence for claims or conclusions. The analysis may include the extent to which the scientific or technological demands of the situation are clearly apparent or to which students must differentiate among components of the situation to clarify the scientific issues as opposed to other, non-scientific issues.

The degree of complexity needed to solve the problem presented: The complexity may range from a single step where students identify the scientific issue, apply a single fact or concept, and present a conclusion to multi-step problems requiring a search for advanced scientific knowledge, complex decision making, information processing and ability to form an argument.

The degree of synthesis needed to answer the question: The synthesis may range from a single piece of evidence where no real construction of justification or argument is required to situations requiring students to apply multiple sources of evidence and compare competing lines of evidence and different explanations to adequately argue a position.

Levels of scientific literacy

The approach to reporting used by the OECD has been defined in previous cycles of PISA and is based on the definition of a number of bands or levels of literacy proficiency. 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 2006 outcomes for science.

For PISA science, 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 scientific literacy that is represented in Figure 15.4 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 15.1.

The information about the items in each band has been used to develop summary descriptions of the kinds of scientific competencies associated with different levels of proficiency. These summary descriptions can then be used to encapsulate typical scientific proficiency of students associated with each level. As a set, the descriptions encapsulate a representation of growth in scientific literacy.

To develop the summary descriptions, growth in scientific competence was first considered separately in relation to items from each of the competencies. Three sets of descriptions were developed. These are presented in following sections, in Figure 15.5. The three sets of descriptions were combined to produce meta-descriptions of six levels of overall scientific literacy, presented here in Figure 15.4.

Table 15.1
Scientific literacy performance band definitions on the PISA scale

Level	Score points on the PISA scale
6	Above 707.9
5	633.3 to 707.9
4	558.7 to 633.3
3	484.1 to 558.7
2	409.5 to 484.1
1	334.9 to 409.5

A clear progression through these levels is apparent in the way in which the individual scientific competencies specified in the PISA scientific literacy framework play out as literacy levels increase.

For example, the competency *identifying scientific issues* is observed to follow a progression through two related dimensions described in combination in Figure 15.5. These dimensions are:

Understanding the methodology of science: At the lowest level students can usually identify variables that are not open to scientific measurement but can do little more than that. Around the middle of the range, levels 3 and 4, there is identification of the independent and dependent variables in an investigation and a developing understanding of both the reason for a control (referent) and the need to account for extraneous variables. Examples of this can be found in SUNSCREENS Q02 and ACID RAIN Q05(1)[2]. As demonstrated by SUNSCREENS Q03, students at these levels can usually identify the question motivating the investigation. At higher levels students are able to view an investigation in its totality and show an awareness of the range of issues that need to be accounted for if meaning is to be ascribed to the outcomes of a testing regime.

Designing an investigation: At the lowest levels students are able to ask questions that elicit relevant information about straightforward scientific issues within familiar contexts. They are able to suggest comparisons to be made given simple cause and effect relationships. Around the middle of the literacy scale students show the capacity to produce simple designs to investigate direct or concrete relationships that are set in relatively familiar contexts. They exhibit an awareness of features that they need to control or account for in their designs. At the highest levels students are able to design ways of investigating questions that involve abstract ideas within the scope of their conceptual knowledge.

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Figure 15.4

Summary descriptions of the six proficiency levels on the science scale

Level	Lower score limit	Percentage of students able to perform tasks at each level or above (OECD average)	What students can typically do
6	707.9	1.3% of students across the OECD can perform tasks at Level 6 on the science scale	At Level 6, students can consistently identify, explain and apply scientific knowledge and *knowledge about science* in a variety of complex life situations. They can link different information sources and explanations and use evidence from those sources to justify decisions. They clearly and consistently demonstrate advanced scientific thinking and reasoning, and they demonstrate willingness to use their scientific understanding in support of solutions to unfamiliar scientific and technological situations. Students at this level can use scientific knowledge and develop arguments in support of recommendations and decisions that centre on personal, social or global situations.
5	633.3	9.0% of students across the OECD can perform tasks at least at Level 5 on the science scale	At Level 5, students can identify the scientific components of many complex life situations, apply both scientific concepts and *knowledge about science* to these situations, and can compare, select and evaluate appropriate scientific evidence for responding to life situations. Students at this level can use well-developed inquiry abilities, link knowledge appropriately and bring critical insights to situations. They can construct explanations based on evidence and arguments based on their critical analysis.
4	558.7	29.3% of students across the OECD can perform tasks at least at Level 4 on the science scale	At Level 4, students can work effectively with situations and issues that may involve explicit phenomena requiring them to make inferences about the role of science or technology. They can select and integrate explanations from different disciplines of science or technology and link those explanations directly to aspects of life situations. Students at this level can reflect on their actions and they can communicate decisions using scientific knowledge and evidence.
3	484.1	56.7% of students across the OECD can perform tasks at least at Level 3 on the science scale	At Level 3, students can identify clearly described scientific issues in a range of contexts. They can select facts and knowledge to explain phenomena and apply simple models or inquiry strategies. Students at this level can interpret and use scientific concepts from different disciplines and can apply them directly. They can develop short statements using facts and make decisions based on scientific knowledge.
2	409.5	80.8% of students across the OECD can perform tasks at least at Level 2 on the science scale	At Level 2, students have adequate scientific knowledge to provide possible explanations in familiar contexts or draw conclusions based on simple investigations. They are capable of direct reasoning and making literal interpretations of the results of scientific inquiry or technological problem solving.
1	334.9	94.8% of students across the OECD can perform tasks at least at Level 1 on the science scale	At Level 1, students have such a limited scientific knowledge that it can only be applied to a few, familiar situations. They can present scientific explanations that are obvious and that follow explicitly from given evidence.

Figure 15.5 [Part 1/2]

Summary descriptions of the six proficiency levels in *identifying scientific issues*

General proficiencies students should have at each level	Tasks a student should be able to do	Examples from released questions
LEVEL 6 1.3% of all students across the OECD area can perform tasks at Level 6 on the *identifying scientific issues* scale.		
Students at this level demonstrate an ability to understand and articulate the complex modelling inherent in the design of an investigation.	▪ Articulate the aspects of a given experimental design that meet the intent of the scientific question being addressed. ▪ Design an investigation to adequately meet the demands of a specific scientific question. ▪ Identify variables that need to be controlled in an investigation and articulate methods to achieve that control.	**ACID RAIN** Question 5
LEVEL 5 8.4% of all students across the OECD area can perform tasks at least at Level 5 on the *identifying scientific issues* scale.		
Students at this level understand the essential elements of a scientific investigation and thus can determine if scientific methods can be applied in a variety of quite complex, and often abstract contexts. Alternatively, by analysing a given experiment can identify the question being investigated and explain how the methodology relates to that question.	▪ Identify the variables to be changed and measured in an investigation of a wide variety of contexts. ▪ Understand the need to control all variables extraneous to an investigation but impinging on it. ▪ Ask a scientific question relevant to a given issue.	
LEVEL 4 28.4% of all students across the OECD area can perform tasks at least at Level 4 on the *identifying scientific issues* scale.		
Students at this level can identify the change and measured variables in an investigation and at least one variable that is being controlled. They can suggest appropriate ways of controlling that variable. The question being investigated in straightforward investigations can be articulated.	▪ Distinguish the control against which experimental results are to be compared. ▪ Design investigations in which the elements involve straightforward relationships and lack appreciable abstractness. ▪ Show an awareness of the effects of uncontrolled variables and attempt to take this into account in investigations.	**SUNSCREENS** Questions 2 and 4 **CLOTHES** Question 1

...

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Figure 15.5 [Part 2/2]

Summary descriptions of the six proficiency levels in *identifying scientific issues*

General proficiencies students should have at each level	Tasks a student should be able to do	Examples from released questions
LEVEL 3 56.7% of all students across the OECD area can perform tasks at least at Level 3 on the *identifying scientific issues* scale.		
Students at this level are able to make judgements about whether an issue is open to scientific measurement and, consequently, to scientific investigation. Given a description of an investigation can identify the change and measured variables.	▪ Identify the quantities able to be scientifically measured in an investigation. ▪ Distinguish between the change and measured variables in simple experiments. ▪ Recognise when comparisons are being made between two tests (but are unable to articulate the purpose of a control).	**ACID RAIN** Question 5 (Partial) **SUNSCREENS** Question 3
LEVEL 2 81.3% of all students across the OECD area can perform tasks at least at level 2 on the *identifying scientific issues* scale.		
Students at this level can determine if scientific measurement can be applied to a given variable in an investigation. They can recognise the variable being manipulated (changed) by the investigator. Students can appreciate the relationship between a simple model and the phenomenon it is modelling. In researching topics students can select appropriate key words for a search.	▪ Identify a relevant feature being modelled in an investigation. ▪ Show an understanding of what can and cannot be measured by scientific instruments. ▪ Select the most appropriate stated aims for an experiment from a given selection. ▪ Recognise what is being changed (the cause) in an experiment. ▪ Select a best set of Internet search words on a topic from several given sets.	**GENETICALLY MODIFIED CROPS** Question 3
LEVEL 1 94.9% of all students across the OECD area can perform tasks at least at Level 1 on the *identifying scientific issues* scale.		
Students at this level can suggest appropriate sources of information on scientific topics. They can identify a quantity that is undergoing variation in an experiment. In specific contexts they can recognise whether that variable can be measured using familiar measuring tools or not.	▪ Select some appropriate sources from a given number of sources of potential information on a scientific topic. ▪ Identify a quantity that is undergoing change, given a specific but simple scenario. ▪ Recognise when a device can be used to measure a variable (within the scope of the student's familiarity with measuring devices).	

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Progression in the *explaining phenomena scientifically* competency can be seen along three dimensions. Descriptions applicable to the various levels can be found in Figure 15.6.

Breadth and depth of scientific knowledge: At the lowest levels students can recall singular scientific facts either learned in a school environment or experienced in daily life in giving simple explanations. Examples of this can be found in CLOTHES Q02 and MARY MONTAGUE Q02. Around the middle of the scale students are able to apply several related pieces of information to an explanation of a phenomenon. In MARY MONTAGUE Q04 students were required to bring knowledge of vaccination, immunity systems and differential resistance in human populations to the question. The knowledge utilised is distinguishable from that of lower levels of literacy by its breadth and the inclusion of an abstract concept where applicable. At the highest levels students can draw upon a broad range of abstract scientific concepts in developing explanations of a phenomenon such as in GREENHOUSE Q05.

Figure 15.6 [Part 1/2]

Summary descriptions of the six proficiency levels in *explaining phenomena scientifically*

General proficiencies students should have at each level	Tasks a student should be able to do	Examples from released questions
LEVEL 6 1.8% of all students across the OECD area can perform tasks at Level 6 on the *explaining phenomena scientifically* scale.		
Students at this level draw on a range of abstract scientific knowledge and concepts and the relationships between these in developing explanations of processes within systems.	▪ Demonstrate an understanding of a variety of complex, abstract physical, biological or environmental systems. ▪ In explaining processes, articulate the relationships between a number of discrete elements or concepts.	**GREENHOUSE** Question 5
LEVEL 5 9.8% of all students across the OECD area can perform tasks at least at Level 5 on the *explaining phenomena scientifically* scale.		
Students at this level draw on knowledge of two or three scientific concepts and identify the relationship between them in developing an explanation of a contextual phenomenon.	▪ Take a scenario, identify its major component features, whether conceptual or factual, and use the relationships between these features in providing an explanation of a phenomenon. ▪ Synthesise two or three central scientific ideas in a given context in developing an explanation for, or a prediction of, an outcome.	
LEVEL 4 29.4% of all students across the OECD area can perform tasks at least at Level 4 on the *explaining phenomena scientifically* scale.		
Students at this level have an understanding of scientific ideas, including scientific models, with a significant level of abstraction. They can apply a general, scientific concept containing such ideas in the development of an explanation of a phenomenon.	▪ Understand a number of abstract scientific models and can select an appropriate one from which to draw inferences in explaining a phenomenon in a specific context (*e.g.* the particle model, planetary models, models of biological systems). ▪ Link two or more pieces of specific knowledge, including from an abstract source in an explanation (*e.g.* increased exercise leads to increased metabolism in muscle cells, this in turn requires an increased exchange of gases in the blood supply which is achieved by an increased rate of breathing).	**PHYSICAL EXERCISE** Question 5

...

Figure 15.6 [Part 2/2]

Summary descriptions of the six proficiency levels in *explaining phenomena scientifically*

General proficiencies students should have at each level	Tasks a student should be able to do	Examples from released questions
LEVEL 3 56.4% of all students across the OECD area can perform tasks at least at Level 3 on the *explaining phenomena scientifically* scale.		
Students at this level can apply one or more concrete or tangible scientific ideas/concepts in the development of an explanation of a phenomenon. This is enhanced when there are specific cues given or options available from which to choose. When developing an explanation, cause and effect relationships are recognised and simple, explicit scientific models may be drawn upon.	▪ Understand the central feature(s) of a scientific system and, in concrete terms, can predict outcomes from changes in that system (*e.g.* the effect of a weakening of the immune system in a human). ▪ In a simple and clearly defined context, recall several relevant, tangible facts and apply these in developing an explanation of the phenomenon.	**MARY MONTAGU** Question 4 **ACID RAIN** Question 2 **PHYSICAL EXERCISE** Question 1
LEVEL 2 80.4% of all students across the OECD area can perform tasks at least at Level 2 on the *explaining phenomena scientifically* scale.		
Students at this level can recall an appropriate, tangible, scientific fact applicable in a simple and straightforward context and can use it to explain or predict an outcome.	▪ Given a specific outcome in a simple context, indicate, in a number of cases and with appropriate cues the scientific fact or process that has caused that outcome (*e.g.* water expands when it freezes and opens cracks in rocks, land containing marine fossils was once under the sea). ▪ Recall specific scientific facts with general currency in the public domain (*e.g.* vaccination provides protection against viruses that cause disease).	**GRAND CANYON** Question 3 **MARY MONTAGU** Questions 2 and 3 **GRAND CANYON** Question 5
LEVEL 1 94.6% of all students across the OECD area can perform tasks at least at Level 1 on the *explaining phenomena scientifically* scale.		
Students at this level can recognise simple cause and effect relationships given relevant cues. The knowledge drawn upon is a singular scientific fact that is drawn from experience or has widespread popular currency.	▪ Choose a suitable response from among several responses, given the context is a simple one and that recall of a single scientific fact is involved (*e.g.* ammeters are used to measure electric current). ▪ Given sufficient cues, recognise simple cause and effect relationships (*e.g.* Do muscles get an increased flow of blood during exercise? Yes or No).	**PHYSICAL EXERCISE** Question 3 **CLOTHES** Question 2

System complexity: At the lowest levels students are able to deal with very simple contexts involving cause and effect relationships as illustrated by PHYSICAL EXERCISE Q03 where the effect of increased exercise is an increase in the flow of blood. Those in the middle ranges of the scientific literacy scale are beginning to view phenomena from a system viewpoint, increasingly extending and recognising the relationships that bear on the phenomenon. The models they understand and use in developing explanations start to deal with abstract scientific ideas and a degree of complexity. PHYSICAL EXERCISE Q05 involves students in drawing on knowledge of the human respiratory system. At the highest levels students show a capacity to develop explanations for contexts with a high degree of complexity involving abstract ideas and sub-systems drawn from a variety of scientific disciplines.

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Synthesis: The ability to bring together relevant concepts and to understand the relationships between them in constructing an explanation of a phenomenon is another dimension that shows progression over the levels of scientific literacy. At level two this is demonstrated in ACID RAIN Q03 where the requirement was to recognise that water would turn to ice below 0°C, know that water expands as it turns to ice and that there is a relationship between that expansion and the breaking down of rock. This involves the synthesis of concrete facts and ideas. By level 4 a progression in this dimension can be seen in PHYSICAL EXERCISE Q05. There, the requirement is to bring into relationship abstract ideas about the need of the muscles of the body for an increased rate of gaseous exchange in the lungs during physical exercise.

Progression in two dimensions is evident in the *using scientific evidence* competency. Descriptions relating to this progression in scientific literacy in this area can be found in Figure 15.5.

Complexity of the data used: At a low level of literacy the student can make comparisons between rows in a simple table or make a conclusion from a simple change in a single variable. A level 2 example of this can be found in GRAND CANYON Q03 where the requirement was to recognise that the loss of gas in a chemical reaction results in a loss of mass for the products left behind. Around the middle of the literacy scale students can utilise data presented in the form of line graphs in making inferences, make simple comparisons between graphs and describe patterns in increasingly complex tables of data. Examples of this can be found in GREENHOUSE Q03 and Q04(1). At higher levels students are able to describe patterns in complex data, summarise that data and suggest explanations for the patterns.

Comparative skills and critical abilities applied to conclusions: Given options or clues, students at the lower levels of this competency can identify a conclusion that is supported by a simple data set. At levels around the middle of the scale students can make judgements about the merit of a conclusion by identifying evidence that is consistent with the conclusion and evidence that does not support it. At the highest levels students can comment on whether evidence is consistent with a given hypothesis and describe the limitations that are inherent in conclusions.

Interpreting the scientific literacy levels

The proficiency levels defined and described in the preceding sections 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 scientific 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 science, 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 15.1 .

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 do we assign individual students to the levels, and what meaning do we ascribe to being at a level? In the context of the OECD reporting of PISA 2000 results, a common sense 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. If we could 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 common-sense interpretation of being at a level is 0.62. Students are placed on the scale at the point where they have a 62% chance of correctly answering test questions located at the same point.

The same meaning has been applied in the reporting of PISA 2000, 2003 and 2006 results. Such an approach makes it possible to summarise aspects of student proficiency by describing the things related to PISA scientific literacy that students can be expected to do at different locations on the scale.

Figure 15.7 [Part 1/2]

Summary descriptions of the six proficiency levels in *using scientific evidence*

General proficiencies students should have at each level	Tasks a student should be able to do	Examples from released questions
LEVEL 6 2.4% of all students across the OECD area can perform tasks at Level 6 on the *using scientific evidence* scale.		
Students at this level demonstrate an ability to compare and differentiate among competing explanations by examining supporting evidence. They can formulate arguments by synthesising evidence from multiple sources.	▪ Recognise that alternative hypotheses can be formed from the same set of evidence. ▪ Test competing hypotheses against available evidence. ▪ Construct a logical argument for an hypothesis by using data from a number of sources.	
LEVEL 5 11.8% of all students across the OECD area can perform tasks at Level 5 on the *using scientific evidence* scale.		
Students at this level are able to interpret data from related datasets presented in various formats. They can identify and explain differences and similarities in the datasets and draw conclusions based on the combined evidence presented in those datasets.	▪ Compare and discuss the characteristics of different datasets graphed on the one set of axes. ▪ Recognise and discuss relationships between datasets (graphical and otherwise) in which the measured variable differs. ▪ Based on an analysis of the sufficiency of the data, make judgements about the validity of conclusions.	**GREENHOUSE** Question 4
LEVEL 4 31.6% of all students across the OECD area can perform tasks at Level 4 on the *using scientific evidence* scale.		
Students at this level can interpret a dataset expressed in a number of formats, such as tabular, graphic and diagrammatic, by summarising the data and explaining relevant patterns. They can use the data to draw relevant conclusions. Students can also determine whether the data support assertions about a phenomenon.	▪ Locate relevant parts of graphs and compare these in response to specific questions. ▪ Understand how to use a control in analysing the results of an investigation and developing a conclusion. ▪ Interpret a table that contains two measured variables and suggest credible relationships between those variables. ▪ Identify the characteristics of a straightforward technical device by reference to diagrammatic representations and general scientific concepts and thus form conclusions about its method of operation.	**SUNSCREENS** Question 5 **GREENHOUSE** Question 4 (Partial)

...

Figure 15.7 [Part 2/2]
Summary descriptions of the six proficiency levels in *using scientific evidence*

General proficiencies students should have at each level	Tasks a student should be able to do	Examples from released questions
LEVEL 3 56.3% of all students across the OECD area can perform tasks at Level 3 on the *using scientific evidence* scale.		
Students at this level are able to select a piece of relevant information from data in answering a question or in providing support for or against a given conclusion. They can draw a conclusion from an uncomplicated or simple pattern in a dataset. Students can also determine, in simple cases, if enough information is present to support a given conclusion.	▪ Given a specific question, locate relevant scientific information in a body of text. ▪ Given specific evidence/data, choose between appropriate and inappropriate conclusions. ▪ Apply a simple set of criteria in a given context in order to draw a conclusion or make a prediction about an outcome. ▪ Given a set of functions, determine if they are applicable to a specific machine.	**GREENHOUSE** Question 3
LEVEL 2 78.1% of all students across the OECD area can perform tasks at Level 2 on the *using scientific evidence* scale.		
Students at this level are able to recognise the general features of a graph if they are given appropriate cues and can point to an obvious feature in a graph or simple table in support of a given statement. They are able to recognise if a set of given characteristics apply to the function of everyday artifacts in making choices about their use.	▪ Compare two columns in a simple table of measurements and indicate differences. ▪ State a trend in a set of measurements or simple line or bar graph. ▪ Given a common artifact can determine some characteristics or properties pertaining to the artifact from among a list of properties.	**ACID RAIN** Question 3
LEVEL 1 92.1% of all students across the OECD area can perform tasks at Level 1 on the *using scientific evidence* scale.		
In response to a question, students at this level can extract information from a fact sheet or diagram pertinent to a common context. They can extract information from bar graphs where the requirement is simple comparisons of bar heights. In common, experienced contexts students at this level can attribute an effect to a cause.	▪ In response to a specific question pertaining to a bar graph, make comparisons of the height of bars and give meaning to the difference observed. ▪ Given variation in a natural phenomenon can, in some cases, indicate an appropriate cause (*e.g.* fluctuations in the output of wind turbines may be attributed to changes in wind strength).	

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Notes

1. While strictly speaking the scales based on aspects of reading are sub-scales of the combined reading literacy scale, for simplicity they are mostly referred to as 'scales' rather than 'sub-scales' in this report.

2. Examples referred to are reproduced in Volume 1 of *PISA 2006: Science Competencies for Tomorrow's World*.

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 and economies

ARG	Argentina
AZE	Azerbaijan
BGR	Bulgaria
BRA	Brazil
CHL	Chile
COL	Colombia
EST	Estonia
HKG	Hong Kong-China
HRV	Croatia
IDN	Indonesia
JOR	Jordan
KGZ	Kyrgyztan
LIE	Liechtenstein
LTU	Lithuania
LVA	Latvia
LVL	Latvia (Latvian Community)
LVR	Latvia (Russian Community)
MAC	Macao-China
MNE	Montenegro
QAT	Qatar
ROU	Romania
RUS	Russian Federation
SRB	Serbia
SVN	Slovenia
TAP	Chinese Taipei
THA	Thailand
TUN	Tunisia
URY	Uruguay

References

Adams, R.J., Wilson, M. & **Wang, W.C.** (1997), The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, No. 21, pp. 1-23.

Adams, R.J., Wilson, M. R. & **Wu, M.L.** (1997), Multilevel item response models: An approach to errors in variables regression, *Journal of Educational and Behavioural Statistics,* No. 22 (1), pp. 46-75.

Adams, R.J. & **Wu, M.L.** (2002), *PISA 2000 Technical Report,* OECD, Paris.

Bollen, K.A. & **Long, S.J.** (1993) (eds.), *Testing Structural Equation Models*, Newbury Park: London.

Beaton, A.E. (1987), Implementing the new design: The NAEP 1983-84 technical report (Rep. No. 15-TR-20), Princeton, NJ: Educational Testing Service.

Buchmann, C. (2000), Family structure, parental perceptions and child labor in Kenya: What factors determine who is enrolled in school? *Soc. Forces,* No. 78, pp. 1349-79.

Buchmann, C. (2002), Measuring Family Background in International Studies of Education: Conceptual Issues and Methodological Challenges, in Porter, A.C. and Gamoran, A. (eds.). *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 150-97), Washington, DC: National Academy Press.

Creemers, B.P.M. (1994), *The Effective Classroom*, London: Cassell.

Cochran, W.G. (1977), *Sampling techniques,* third edition, New York, NY: John Wiley and Sons.

Ganzeboom, H.B.G., de Graaf, P.M. & **Treiman, D.J.** (1992), A standard international socio-economic index of occupational status, *Social Science Research*, No. 21, pp. 1-56.

Ganzeboom H.B. & **Treiman, D.J.** (1996), Internationally comparable measures of occupational status for the 1988 international standard classification of occupations, *Social Science Research*, No. 25, pp. 201-239.

Grisay, A. (2003), Translation procedures in OECD/PISA 2000 international assessment, *Language Testing,* No. 20 (2), pp. 225-240.

Hambleton, R.K., Swaminathan, H. & **Rogers, H.J.** (1991), *Fundamentals of item response theory*, Newbury Park, London, New Delhi: SAGE Publications.

Hambleton, R.K., Merenda, P.F. & **Spielberger, C.D.** (2005), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment,* IEA Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey.

Harkness, J.A., Van de Vijver, F.J.R. & **Mohler, P.Ph** (2003), *Cross-Cultural Survey Methods,* Wiley-Interscience, John Wiley & Sons, Inc., Hoboken, New Jersey.

Harvey-Beavis, A. (2002), Student and School Questionnaire Development, in R.J. Adams and M.L. Wu (eds.), *PISA 2000 Technical Report,* (pp. 33-38), OECD, Paris.

International Labour Organisation (ILO) (1990), *International Standard Classification of Occupations: ISCO-88.* Geneva: International Labour Office.

Jöreskog, K.G. & **Sörbom, Dag** (1993), *LISREL 8 User's Reference Guide*, Chicago: SSI.

Judkins, D.R. (1990), Fay's Method of Variance Estimation, *Journal of Official Statistics*, No. 6 (3), pp. 223-239.

Kaplan, D. (2000), *Structural equation modeling: Foundation and extensions*, Thousand Oaks: SAGE Publications.

Keyfitz, N. (1951), Sampling with probabilities proportionate to science: Adjustment for changes in probabilities, *Journal of the American Statistical Association,* No. 46, American Statistical Association, Alexandria, pp. 105-109.

Kish, L. (1992), Weighting for Unequal, *Pi. Journal of Official Statistics*, No. 8 (2), pp. 183-200.

LISREL (1993), K.G. Jöreskog & D. Sörbom, [computer software], Lincolnwood, IL: Scientific Software International, Inc.

Lohr, S.L. (1999), *Sampling: Design and Analysis*, Duxberry: Pacific Grove.

Macaskill, G., Adams, R.J. & **Wu, M.L.** (1998), Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scale, in M. Martin and D.L. Kelly, Editors, *Third International Mathematics and Science Study, technical report Volume 3: Implementation and analysis*, Boston College, Chestnut Hill, MA.

Masters, G.N. & **Wright, B.D.** (1997), The Partial Credit Model, in W.J. van der Linden, & R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory* (pp. 101-122), New York/Berlin/Heidelberg: Springer.

Mislevy, R.J. (1991), Randomization-based inference about latent variables from complex samples, *Psychometrika,* No. 56, pp. 177-196.

Mislevy, R.J., Beaton, A., Kaplan, B.A. & **Sheehan, K.** (1992), Estimating population characteristics from sparse matrix samples of item responses, *Journal of Educational Measurement,* No. 29 (2), pp. 133-161.

Mislevy, R.J. & **Sheehan, K.M.** (1987), Marginal estimation procedures, in Beaton, A.E., Editor, 1987. *The NAEP 1983-84 technical report*, National Assessment of Educational Progress, Educational Testing Service, Princeton, pp. 293-360.

Mislevy, R.J. & **Sheehan, K.M.** (1989), Information matrices in latent-variable models, *Journal of Educational Statistics*, No. 14, pp. 335-350.

Mislevy, R.J. & **Sheehan, K.M.** (1989), The role of collateral information about examinees in item parameter estimation, *Psychometrika*, No. 54, pp. 661-679.

Monseur, C. & **Berezner, A.** (2007), The Computation of Equating Errors in International Surveys in Education, *Journal of Applied Measurement,* No. 8 (3), 2007, pp. 323-335.

Monseur, C. (2005), An exploratory alternative approach for student non response weight adjustment, *Studies in Educational Evaluation*, No. 31 (2-3), pp. 129-144.

Muthen, B. & **L. Muthen** (1998), [computer software], *Mplus* Los Angeles, CA: Muthen & Muthen.

Muthen, B., du Toit, S.H.C. & **Spisic, D.** (1997), *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*, unpublished manuscript.

OECD (1999), *Classifying Educational Programmes. Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2003), *Literacy Skills for the World of Tomorrow: Further results from PISA 2000*, OECD, Paris.

OECD (2004), *Learning for Tomorrow's World – First Results from PISA 2003*, OECD, Paris.

OECD (2005), *Technical Report for the OECD Programme for International Student Assessment 2003*, OECD, Paris.

OECD (2006), *Assessing Scientific, Reading and Mathematical Literacy: A framework for PISA 2006,* OECD, Paris.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World*, OECD, Paris.

PISA Consortium (2006), *PISA 2006 Main Study Data Management Manual*, *https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2006_data_management_manual.pdf*

Rasch, G. (1960), Probabilistic models for some intelligence and attainment tests, Copenhagen: Nielsen & Lydiche.

Routitski A. & **Berezner, A.** (2006), Issues influencing the validity of cross-national comparisons of student performance. Data Entry Quality and Parameter Estimation. Paper presented at the Annual Meeting of the American Educational Research Association (AERA) in San Francisco, 7-11 April, *https://mypisa.acer.edu.au/images/mypisadoc/aera06routitsky_berezner.pdf*

Rust, K. (1985), Variance Estimation for Complex Estimators in Sample Surveys, *Journal of Official Statistics*, No. 1, pp. 381-397.

Rust, K.F. & **Rao, J.N.K.** (1996), Variance Estimation for Complex Surveys Using Replication Techniques, *Survey Methods in Medical Research*, No. 5, pp. 283-310.

Shao, J. (1996), Resampling Methods in Sample Surveys (with Discussion), *Statistics*, No. 27, pp. 203-254.

Särndal, C.-E., Swensson, B. & **Wretman, J.** (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

SAS® CALIS (1992), W. Hartmann [computer software], Cary, NC: SAS Institute Inc.

Scheerens, J. (1990), School effectiveness and the development of process indicators of school functioning, *School effectiveness and school improvement,* No. 1, pp. 61-80.

Scheerens, J. & **Bosker, R.J.** (1997), *The Foundations of School Effectiveness*, Oxford: Pergamon.

Schulz, W. (2002), Constructing and Validating the Questionnaire composites, in R.J. Adams and M.L. Wu (eds.), *PISA 2000 Technical Report*, OECD, Paris.

Schulz, W. (2004), Mapping Student Scores to Item Responses, in W. Schulz and H. Sibberns (eds.), *IEA Civic Education Study, Technical Report* (pp. 127-132), Amsterdam: IEA.

Schulz, W. (2006a), *Testing Parameter Invariance for Questionnaire Indices using Confirmatory Factor Analysis and Item Response Theory,* Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Schulz, W. (2006b), *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and PISA 2003*, Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Thorndike, R.L. (1973), *Reading comprehension in fifteen countries,* New York, Wiley: and Stockholm: Almqvist & Wiksell.

Travers, K.J. & **Westbury, I.** (1989), *The IEA Study of Mathematics I: Analysis of Mathematics Curricula*, Oxford: Pergamon Press.

Travers, K.J., Garden R.A. & **Rosier, M.** (1989), Introduction to the Study, in Robitaille, D. A. and Garden, R. A. (eds), *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics Curricula,* Oxford: Pergamon Press.

Verhelst, N. (2002), Coder and Marker Reliabilaity Studies, in R.J. Adams & M.L. Wu (eds.), *PISA 2000 Technical Report*. OECD, Paris.

Walberg, H.J. (1984), Improving the productivity of American schools, *Educational Leadership,* No. 41, pp. 19-27.

Walberg, H. (1986), Synthesis of research on teaching, in M. Wittrock (ed.), *Handbook of research on teaching* (pp. 214-229), New York: Macmillan.

Walker, M. (2006), *The choice of Likert or dichotomous items to measure attitudes across culturally distinct countries in international comparative educational research*. Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Walker, M. (2007), Ameliorating Culturally-Based Extreme Response Tendencies To Attitude items, *Journal of Applied Measurement,* No. 8, pp. 267-278.

Warm, T.A. (1989), Weighted Likelihood Estimation of Ability in Item Response Theory, *Psychometrika*, No. 54 (3), pp. 427-450.

Westat (2007), *WesVar® 5.1* Computer software and manual, Rockville, MD: Author (also see *http://www.westat.com/wesvar/*).

Wilson, M. (1994), Comparing Attitude Across Different Cultures: Two Quantitative Approaches to Construct Validity, in M. Wilson (ed.), *Objective measurement II: Theory into practice* (pp. 271-292), Norwood, NJ: Ablex.

Wolter, K.M. (2007), *Introduction to Variance Estimation*. Second edition, Springer: New York.

Wu, M.L., Adams, R.J. & **Wilson, M.R.** (1997), *ConQuest®: Multi-Aspect Test Software* [computer program manual], Camberwell, Vic.: Australian Council for Educational Research.

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
CBAS	Computer Based Assessment of Science
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CITO	National Institute for Educational Measurement, The Netherlands
CIVED	Civic Education Study
DIF	Differential Item Functioning
ENR	Enrolment of 15-year-olds
ESCS	PISA Index of Economic, Social and Cultural Status
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
SE	Standard Error
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

Table of contents

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

LIST OF BOXES

LIST OF FIGURES

LIST OF TABLES

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009