PISA – The Programme for International Student Assessment – An Overview

An overview of the PISA project is given in order to set a context for a consideration of the ways in which this large international assessment programme uses scientific measurement principles and techniques. The overview describes the goals of the project, its major organisational features (its participants, its procedures and organisational structure, its main implementation steps) and its major technical components (test design, test development, sampling, assessment operations, data scaling and analysis, and reporting). The ways in which measurement principles are used are interwoven with the discussion of the relevant technical components. The paper concludes with a detailed discussion of the construction of the described proficiency scales used to report outcomes for PISA mathematics.

What is PISA?

The Programme for International Student Assessment (PISA) was designed and developed by the Organisation for Economic Co-operation and Development (OECD) in the late 1990s as an ongoing, periodic international comparative study of certain student characteristics and proficiencies. It is designed to generate indicators of aspects of educational performance, so that the participating countries have access to high-quality and reliable measures of *outcomes* of aspects of their educational systems. PISA is therefore not primarily a research project, though the data generated may be of great interest to researchers. PISA is managed and directed cooperatively by the OECD member countries, and in cooperation with a large and increasing number of non-member countries, referred to as 'partner' countries. The OECD administers the project through a small Secretariat based in Paris. For each survey cycle, the OECD appoints an external contractor to implement the project following an open competitive tendering process.

PISA surveys take place every three years. The first survey took place during 2000, and the second in 2003. The third survey is occurring during 2006, and the next will occur in 2009. Scientific sampling procedures are used to determine which schools and which students will be included in each survey. A common set of assessment and survey instruments are used in each participating country under common and controlled conditions that enable comparisons to be made based on the resulting data. Analytic techniques are used that enable comparisons within and among participating countries, and across survey cycles.

PISA is an age-based survey, assessing 15-year-old students in school in grade 7 or higher. These students are approaching the end of compulsory schooling in most participating countries, which makes this a suitable age-group at which to target an assessment of the extent to which students are prepared for the daily challenges of modern societies.

To do this, PISA takes a 'literacy' perspective that focuses on the extent to which students can use the knowledge and skills they have learned and practised at school when confronted with situations and challenges for which that knowledge may be relevant. That is, it assesses the extent to which students can use their reading skills to understand and interpret various kinds of written material that they are likely to meet as they negotiate their daily lives; the extent to which students can use their mathematical knowledge and skills to solve various kinds of mathematics-related challenges and problems they are likely to meet; and the extent to which students can use their scientific knowledge and skills to understand, interpret and resolve various kinds of scientific situations and challenges.

PISA also allows for the assessment of additional cross-curricular competencies from time to time as participating countries see fit. For example, in the 2003 survey cycle, an assessment of general problem-solving competencies was included. Further, the PISA survey collects information from students on various aspects of their home, family and school background; and information from schools about various aspects of organisation and educational provision in schools. This information is collected to facilitate a detailed study of factors within and between countries that are associated with varying levels of reading, mathematical and scientific literacy among the 15-year-old students of each country. The resulting analyses will be of interest to policy makers in participating countries seeking to better understand the relationships between performance and a variety of background factors, the connections to various national education policy settings, and the responsiveness of outcomes to changes in policy settings. The data will also be of real interest to researchers seeking to better understand the factors influencing educational outcomes.

Features of PISA – Organisational

The PISA project can be thought of as operating at a number of levels. The different levels and key participants in the project are illustrated in Figure 1.

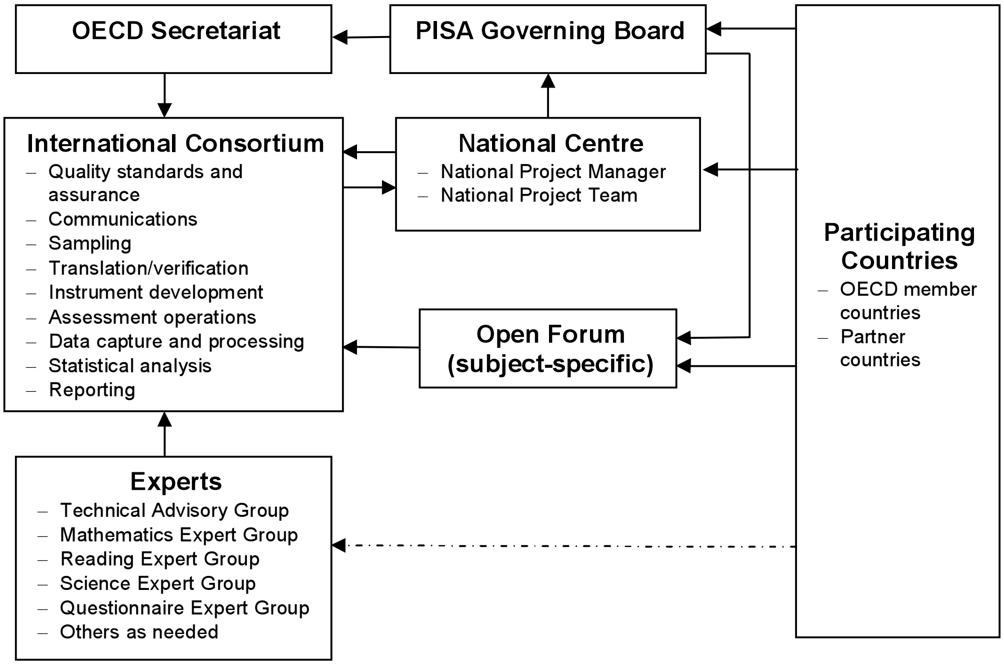

Figure 1. Key Participants in the PISA Project

The PISA project is administered by a Secretariat within the OECD in Paris. The project is overseen and policy parameters are set by the PISA Governing Board (PGB), which is an OECD committee comprising delegates and observers from the participating countries. It meets about twice each year. PISA therefore is directed through a collaborative process involving high level educational administrators from participating countries.

Within each participating country, a national centre is established. A National Project Manager (NPM) is appointed to coordinate all activities at the national level. Typically, the NPM works closely with the country's PGB member to establish a national perspective on policy matters, on matters related to project implementation, and on the analysis and reporting of outcomes that may be of particular relevance to the country. Generally the national centre has a small team working on project development, implementation and reporting at the national level.

The project is implemented internationally by a contractor appointed by the OECD to carry out this work. A contractor is selected for each assessment cycle following an open competitive tendering process. An international consortium led by the Australian Council for Educational Research (ACER) has been the successful contractor for each of the first three PISA survey cycles. The contractor is responsible for implementing all aspects of the assessment, under the strict guidance of the Secretariat. This includes a wide range of activities, several of which are listed here.

- developing quality standards in relation to all aspects of the project,
- developing procedures to ensure that those quality standards are properly met by all participants, and mechanisms for monitoring the quality of project implementation,
- maintaining open and effective communications among all project participants,
- working closely with National Centre personnel to gather national input to matters related to project development and implementation,
- developing the assessment frameworks,
- developing all assessment and survey instruments,
- developing and implementing sampling plans,
- developing operational procedures for test administration and all related documentation,
- training key national centre staff in the requirements for implementing the study,
- developing all data capture procedures,
- capturing and processing data from the assessments,
- analysing the results and preparing material to assist the OECD in producing the reports they require,
- assisting the Secretariat with review of procedures and planning improvements for future survey cycles.

To carry out all of these tasks, the contractor is required firstly to put together a large team of people with expertise collectively in a variety of areas. Some of this expertise resides in the staff available through its consortium partner organisations. In addition, the consortium works with several key groups and individuals: a number of expert groups comprising internationally recognised experts in areas such as the particular cognitive assessment domains, translation experts, technical specialists in sampling, statistical analysis and development of questionnaires. The contractor engages other expertise when required.

Most importantly, the contractor works very closely with the personnel engaged at the national level. The connection between National Project Managers and the international consortium is critical to the success of the project. Each depends on the other to ensure successful implementation of the project. National centres provide the consortium with information about conditions and constraints operating in the country, with feedback regarding the various project elements that are being developed, and with important advice on how the project can best be implemented in that country. The contractor provides national centres with information about project requirements, draft materials for national consideration and feedback, training and materials to facilitate project implementation.

Other consultative mechanisms have been used from time to time, such as the Science Forum for PISA 2006. This is an open forum that provides the opportunity for participating countries to nominate national experts who can directly represent the interests and views of the country in considering certain detailed technical aspects of the project. It allows for a wider base of expert input than is possible through the international contractor's expert groups. In the case of the Science Forum, this group considered priorities and issues at the time the science framework was being conceptualised, and it has provided important input in the development of survey material related to the assessment of science and the assessment of student attitudes to science. Delegates to the forum are nominated by each country's PGB member. A similar Mathematics Forum operated in PISA 2003, and a Questionnaire Forum has been convened from time to time.

Who participates?

Fifty-seven countries participate in the current PISA assessment cycle, including all 30 OECD countries, and 27 other national entities, referred to as 'partner countries'. They are listed in Table 1.

OECD COUNTRIES

Australia
Austria
Belgium
Canada
Czech Republic
Denmark
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Japan
Korea
Luxembourg
Mexico
Netherlands
New Zealand
Norway
Poland
Portugal
Slovak Republic
Spain
Sweden
Switzerland
Turkey
UK
USA

PARTNER COUNTRIES

Argentina
Azerbaijan
Brazil
Bulgaria
Chile
China[1] SAR – Hong Kong
China SAR – Macau
Colombia
Croatia
Estonia
Indonesia
Israel
Jordan
Kyrgyzstan
Latvia
Liechtenstein
Lithuania
Qatar
Romania
Russian Federation
Serbia and Montenegro
Slovenia
Taiwan
Thailand
Tunisia
Uruguay

Table 1. Countries participating in PISA 2006

For these countries, over 40 different languages are involved.

Other personnel involved in the PISA 2003 project as PGB members, national project managers, staff of the OECD Secretariat, staff of the international consortium,

[1] China does not participate as a single country, but two of its special administrative regions (Hong Kong and Macao) participate in PISA 2006 as if each were a separate country.

members of expert groups, or other consultants, are listed in the OECD's international report (OECD 2004, p. 474).

Management and implementation of PISA

The PISA main survey occurs each three years, and the following discussion presents an overview of the timeline and process for project implementation, based on the current PISA 2006 assessment. For the PISA 2006 survey cycle, work commenced toward the end of 2003 (before main study testing had been completed for the previous survey cycle, and well before any main study data from that earlier cycle had been processed and analysed). The first tasks related to development of frameworks, and to establishment of the required expertise within the international consortium and among the consortium's expert groups.

The international consortium was invited by the OECD Secretariat to facilitate a process of review and revision of the science framework during the latter part of 2003. The Science Forum took the initial steps in reviewing the Science framework, and then the consortium's science expert group was formed to carry this work to final conclusion. The expert groups comprise recognised and respected academics and practitioners in the relevant domain areas, and their role is to guide the intellectual development work, within the policy framework set by the PGB and Secretariat. Participating countries had a number of opportunities to provide input to and feedback on the framework and other documents as they developed. The resulting framework documents would later provide the conceptual basis on which the assessments in each area would be built. Test item development commenced as soon as the directions of the respective frameworks had advanced sufficiently.

Near the beginning of 2004, by which time national centres had been established in most of the participating countries, a meeting of National Project Managers was convened, and preparations for the PISA 2006 survey commenced in earnest. Countries were invited to prepare and submit assessment items for possible use by the consortium. During the next several months, test item development went ahead, coordinated from ACER, and involving test development groups in the organisations that were part of or working closely with the international consortium on this task. Draft test items were developed, they were subjected to various pilot testing activities, they were circulated to national centres for comment, they were reviewed by the relevant domain expert groups, and eventually a selection of items was chosen and finalised for field trial.

Source materials prepared by the international contractor were provided to participating countries in both English and French. This meant that countries had two equivalent source versions from which to build their local language versions of the test and questionnaire instruments. Standard translation procedures involved countries using the two source versions to generate two independent translations in each national language, and to use an expert to reconcile these into a single translated version. The international contractor then arranged for an independent expert verification process to be applied to ensure the linguistic quality of all national versions.

During 2005 an extensive field trial took place in all participating countries. The purpose of the field trial was two-fold. First, all the procedures needed for the main survey were developed and tested, and countries had the direct experience of putting into place all of these procedures across a large number of schools. This included sampling of schools, negotiating school participation, sampling of students, negotiating student participation, translation of test materials into the relevant local languages, recruitment and training of local staff in the required test administration procedures, preparation of all test materials ready for use in schools, implementation of the test administration procedures in the sampled schools, collection and coding of student responses, capture of data, cleaning and processing of data, and submission of data to the international contractor for analysis. Second, the test and questionnaire items selected for the field trial were implemented across a very large number of students and schools in each country, generating data that were subsequently used to determine the quality of the test items, and therefore to inform final test and questionnaire item development and selection for the main survey.

Data and information from the field trial were captured and processed by the international contractor during 2005. The test and questionnaire items were reviewed and refined in light of those data, and survey instruments were finalised late in 2005 for use in the main survey. Main survey instruments and materials were subsequently dispatched to all participating countries, and the personnel at each national centre prepared for the main survey that took place during 2006. All data captured from the main survey will be analysed during 2007, in preparation for the international release of results that will take place in December 2007.

Features of PISA – Technical

The technical characteristics of the PISA survey involve a number of different elements. The design of the test, and the features incorporated into the test developed for PISA are critical features. Test development makes use of Rasch analysis. The sampling design, including both the school sampling and the student sampling requirements and procedures are a second critical area. Features related to the multi-lingual nature of the test are a further area raising a number of important technical issues. This involves the rules and procedures designed to guarantee the equivalence of the different language versions used within and between participating countries, and taking into account the diverse cultural contexts of those countries. Various operational procedures, including test administration arrangements, data capture and processing and quality assurance mechanisms designed to ensure the generation of comparable data from all countries form another area of technical focus for the project. Then of course the technicalities related to scaling and analysis of the data, and their subsequent reporting, form a further major set of issues. PISA scaling employs models based on Rasch methodologies, and the use of described proficiency scales as a basic tool in reporting PISA outcomes also are derived using Rasch analysis. Each of these technical areas will be briefly discussed in turn in the sections following. Greater detail is provided on each of these areas in the Technical Report for PISA 2003 (OECD 2005).

The highest quality standards within each of these technical areas are defined, monitored and assured through the use of a set of technical standards. These standards have been established by the PISA Governing Board, and they form the backbone of project implementation in each participating country and of quality assurance across the project.

Test design and development

PISA has so far been implemented using pencil-and-paper tests. Students are expected to undertake two hours of testing in the main ‘cognitive’ test that covers the domains of reading, science and mathematics. For the PISA 2006 assessment, a number of test items aimed at exploring student attitudes to science were embedded in the cognitive part of the test. In addition, students complete a short questionnaire designed to gather relevant background data about the student's personal characteristics, opinions, preferences and aspirations, some characteristics of his or her home and family environment, and some characteristics of his or her school environment. This is designed for students to complete in 20-30 minutes. School principals also complete a short questionnaire, about broader aspects of the school context.

The cognitive part of the test must provide suitable coverage of each test domain, and must generate data related to the several constructs laid out in each of the assessment frameworks (OECD 2006). In other words, the development of assessment instruments commences with an explicitly stated set of constructs to be targeted by specific test items. The constructs encompass a range of aspects of subject content within each domain, a range of learning processes relevant to each domain, and a variety of contexts that are used in the presentation of test items to ensure that no particular set of interests and experiences is unfairly over-represented, and so that the wide variety of student experiences in different national contexts is adequately represented in the test. The PISA tests are designed to measure the extent to which students can use the range of knowledge and skills they have acquired at school, as they attempt to solve the kinds of problems they will confront in non-school contexts. The emphasis is not on assessing specific curricular outcomes, but on the application of acquired knowledge in a variety of real-life contexts.

A large volume of test material is therefore developed in each assessment cycle to ensure suitable coverage and balance across the various constructs and aspects of each domain framework, and the material is distributed across a number of test booklets in a rotated test design (a balanced incomplete block design). Each sampled student is randomly assigned one of the test booklets.

Rasch analysis is used during the test development process to check the characteristics of the items developed prior to their finalisation and selection for inclusion in the main survey instruments. In particular, the extensive field trial that takes place in the year preceding the main survey in each assessment cycle generates student response data on all items that are being considered for inclusion in the main survey item pool. ConQuest (Wu, Adams and Wilson, 1999) is used as the main analysis tool. The standard item statistics generated by ConQuest as well as various Rasch fit statistics and diagnostic indicators are used as primary tools in reviewing item performance. These include indices of discrimination and fit to the model, point biserial correlations, the mean ability of students by response category, a check of category ordering for partial credit items and the consistency of this across countries, the expected and observed score curves by gender and by country, the expected and observed item characteristic curves by response category.

The information from these analyses is used as the basis of item selection for the main survey. It is also typically used as the basis for identifying items that need to be revised (for example evidence from an unexpected item performance in an individual country that leads to uncovering of a translation error), or that have characteristics that render them unsuitable for use in the PISA assessment instruments. Information about item difficulty is particularly important in the construction of survey instruments, since it is a requirement that the instruments should contain test items with an acceptable mix of difficulties within each of the relevant framework categories.

PISA test items are presented in several different item formats: multiple-choice, short-answer, and extended response. Multiple-choice items are either standard multiple-choice with a limited number of responses (usually four) from which students are required to select the best answer, or complex multiple-choice presenting several statements for each of which students are required to choose one of two or more possible responses (true/false, correct/incorrect, etc.). Short-answer items include 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 only a word or short phrase. Short response items are similar to closed constructed-response items, but for these a wider range of responses is possible. Open constructed-response items have a much wider range of acceptable responses. They typically require more extensive writing, or showing a calculation, or demonstrating a chain of reasoning, and frequently demand some explanation or justification.

In previous PISA cycles, it has been successfully argued by the test developers, by domain experts, and reinforced by national feedback on items as they were in development, that the range of cognitive processes that can be exposed and tapped is much greater when open formats are used than would be the case if only closed form items such as multiple-choice and short response items were used. The PISA Governing Board has taken the view that the additional costs involved with coding and processing responses from these more open items are justified by the increased power and richness of data derived from using a wider range of test item formats. This view has also been reinforced by research on item format using PISA data (Routitsky and Turner, 2003) that has shown the importance of using a variety of test item formats to cater for the full range of student abilities typically sampled in PISA.

Student responses to more than half of the cognitive test items in the PISA 2006 main survey were able to be processed by computer. The remainder, a total of 80 of the 185 items (that is, 43%) required intervention by a trained coder in order to process student responses.

A common battery of questionnaire items was chosen for the background questionnaires. The purpose of the background questionnaires was to identify social, cultural, economic and educational factors that are associated with student performance. This would make it possible to explore the relationships between student performance outcomes on the cognitive tests and various student-level and school-level factors, and to see how these factors might vary across systems, across countries, and across time.

Sampling in PISA

PISA sampling is carried out in two stages, according to a procedure that is designed to assign all eligible students in each participating country a known probability of being chosen to participate. The international population definition enables construction of a sampling frame that comprises all 15-year-old students in school, in grade 7 or higher. First, schools that contain eligible students are randomly sampled with probability proportional to size. Then 35 eligible students are randomly sampled from within each sampled school. In other words, the students who are sampled for the PISA tests are randomly selected, and truly represent the population of 15-year-old students in school in each participating country.

A minimum of 150 schools are sampled in each country (or all schools if there are fewer than 150 containing eligible students). The target student sample size of 35 per school means that a minimum of 5250 students from each country would be sampled, with the expectation that a minimum of 4500 students would be assessed. If fewer than 35 students are available in a large enough number of schools, then additional schools are sampled to ensure an adequate minimum total student sample size.

Standards are applied to ensure adequate coverage of the eligible population (involving strict rules about which schools and students could be legitimately excluded), adequate accuracy and precision in the estimates derived from the sample (involving strict rules about the required sample size), and adequate response rates for both schools and students (involving strict rules about response rates, including procedures for using replacement schools where needed to reach acceptable school response rates and decision rules about inclusion or exclusion of student data depending on student response rates).

Translation and cultural appropriateness of PISA material

PISA is the largest ever international study of its kind. In PISA 2006, it involved test administration in at least 150 different schools in each of the 57 countries participating, involving 81 different verified national versions of assessment instruments in 42 different languages. In such an enterprise, the need to ensure comparability of the test material across all test administrations is no small matter. The first part of achieving this lies in ensuring that the test materials themselves are appropriate for use across such culturally diverse settings, and that the different versions used in those different settings are equivalent.

The approach to ensuring cultural appropriateness is first to use a wide variety of materials, representing different cultural experiences and contexts, then to process and refine those materials to ensure that different interests are well balanced, and to empirically test that all selected materials work well in all countries. The mechanisms used to ensure that materials developed for use in PISA are culturally appropriate include the following:

- Test materials (questionnaire and test items) are sought from the widest possible range of sources, including seeking national submissions from all participating countries, and test development procedures are used that are overseen by international experts in the relevant field, and conducted by test development experts from a variety of countries and cultural contexts.
- Several opportunities are provided for all participating countries to review and comment on the material under development; and material is also reviewed by panels of international experts in each development area.
- Cognitive laboratory and other pilot activities are conducted with material under development using real students in several different countries.
- A large-scale formal field trial is conducted with students in all participating countries to test the functioning of material under development.
- The results of statistical analysis of field trial data are used to empirically evaluate the material as implemented in all countries and to detect instances of test items behaving differently in different countries.
- All material is revised on the basis of information received from each of these different mechanisms, and only material that is demonstrated to work is selected for use in the main PISA assessment.

- The decision about which material is finally selected is reviewed by National Project Managers, and endorsed by the PISA Governing Board.

In parallel with the development of culturally appropriate material for use in the PISA assessment instruments, source versions of all material are prepared in both English and French as a precursor to the development of equivalent national versions. From these two source versions, national centres in each participating country then prepare their own national versions of the test material, using a tightly controlled process of double independent translation, expert reconciliation of the two versions so produced, and independent international expert verification of the final translated versions.

An extensive field trial is conducted to test the translated materials of each country, and analyses of field trial data are used to empirically evaluate the equivalence of the different language and national versions. Information about characteristics of items as they occur in the different language instruments is obtained from the Rasch analyses mentioned previously. One particularly important output is the 'item by country interaction' data (a form of DIF analysis) that are used to expose any items that behave differently when presented in a particular language or culture. This is essential to ensuring that the test instrument use for the main survey comprises items that together are capable of generating a single international set of item difficulty parameters that be used for measurement of students.

Materials are revised on the basis of the field trial data analyses, and the final selection of material for the main PISA study is chosen to ensure that only fully functioning test items across all national and language versions are selected. At the conclusion of this process, each national version that is produced can be regarded as linguistically and psychometrically equivalent to the source versions, and therefore capable of contributing to the estimation of a single international set of item parameters.

Field operations

The second major way in which comparability of test results across such a diverse range of countries and settings is ensured lies in the standardisation of test administration procedures. An extensive array of procedures has been developed and documented to assist all participating countries to administer PISA test sessions in a way that facilitates the generation of internationally comparable test data.

A National Project Manager's manual describes all procedures to be developed and implemented by each national centre, including involvement in various consultation and review procedures, the implementation of sampling procedures, implementation of all procedures related to the preparation, production and dispatch of test materials, recruitment and training of test administration personnel and oversight of test administration, assistance with implementing quality monitoring procedures, recruitment and training of personnel to code student responses, management of the coding of student responses and the entry of student response data, processes related to the capture and preparation of all PISA data for submission to the international contractor, and subsequent processes related to assisting with the analysis of data and the reporting of results.

Separate manuals cover specific operational procedures related to sampling, translation, test administration, test centre coordination, coding of student responses, data management, and related to the specialised data capture software used in the project. As well as this extensive documentation, the international contractor conducts several meetings of National Project Managers and other key national centre staff, the main purpose of which is to provide information and training related to field operations and all other aspects of project implementation in each country.

Through the extensive field operations documentation, the scheduled training and briefing meetings, and by using regular communication via e-mail and telephone, the international contractor ensures that PISA testing procedures could be consistently applied in all participating countries. To check the extent to which those procedures were in fact applied consistently, a variety of quality-monitoring procedures are implemented, and these are described in the following section.

Monitoring quality

Implementation of the PISA project is built around a set of quality standards that relate to the various aspects of the project. Standards exist in relation to the definition of the target population, sampling, language of testing, preparation of tests and manuals, test administration, print quality of test materials, security of materials, and so on.

For each such area, a plan is established that describes the manner in which the standard is to be met by each participating country, evidence is generated as the project proceeds that shows whether each standard has been met, control processes exist through which national PISA centres continuously ensure the quality of project implementation, and external monitoring procedures are applied that enable early intervention by the international contractor in cases for which it appears a standard may not be met in a particular participating country, or intervention after the fact to determine if the data collected can safely be used for the intended purpose.

Ultimately, an adjudication process is in place through which a final determination is made as to whether or not the data from a participating country are fit for use, and can therefore be included in the international dataset. If they are not deemed to be fit for use, then a country's data can be excluded from some or all of the international reporting of PISA outcomes. The adjudication process involves senior consortium staff, members of the technical advisory group including the sampling referee, and the OECD Secretariat. Recommendations from that group are made to the PISA Governing Board for final decision.

Scaling and analysis of data

The scaling of PISA data rests on a simple assumption that there is some underlying trait or set of underlying traits of interest (these traits are defined in the assessment frameworks across the several test domains) that each form a continuum, or scale; that test items can be developed that embody each underlying trait (by demanding different amounts of that trait); and that the amount of the trait possessed by students can be estimated by observing their responses to the test items. Typically, we think of the continuum as a line, with 'more' of the trait in one direction, and 'less' of the trait in the other direction; that test items can be placed along the line according to the amount of the underlying trait that they demand; and that students can also be located along the line according to the amount of the underlying trait that they possess. The Rasch analysis provides a means for constructing an interval scale for these measures. Figure 2 provides a graphic characterisation of a PISA literacy scale, labelled as the Mathematics scale, but the same idea applies to scales developed in any of the PISA test domains.

PISA uses a Rasch-based form of item response modelling in order to scale the student data to derive the various comparative measures that are produced and reported by the OECD. The model is a ‘mixed co-efficients multinomial logit model’, which is a generalised form of the Rasch model. Essentially it uses student responses to a set of test items to simultaneously derive estimates of the ‘difficulty’ of the test items, and of the ‘ability’ of the sampled students, enabling both test items and students to be located along a continuum that is defined by the underlying trait being measured. Details of the model used and the scaling methods applied can be found in the PISA 2003 Technical Report (OECD 2005).

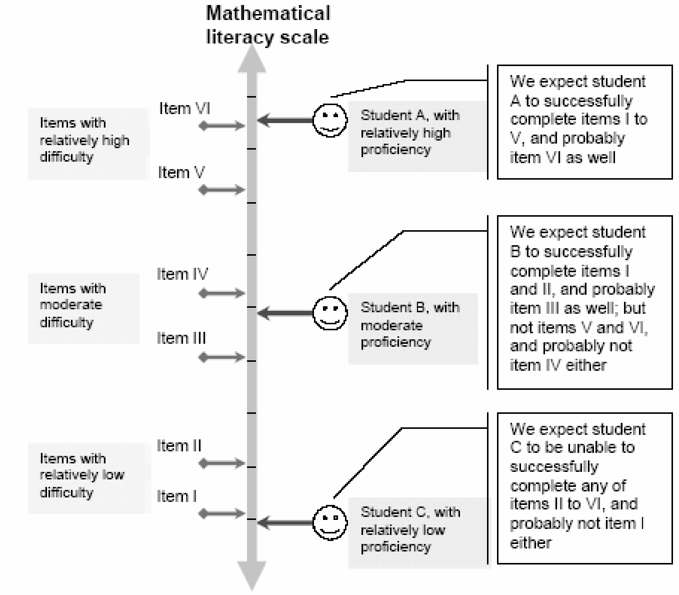
Mathematical literacy scale
Items with relatively high difficulty
Item VI
Item V
Student A, with relatively high proficiency
We expect student A to successfully complete items I to V, and probably item VI as well
Items with moderate difficulty
Item IV
Item III
Student B, with moderate proficiency
We expect student B to successfully complete items I and II, and probably item III as well; but not items V and VI, and probably not item IV either
Items with relatively low difficulty
Item II
Item I
Student C, with relatively low proficiency
We expect student C to be unable to successfully complete any of items II to VI, and probably not item I either

Figure 2. A characterisation of a PISA literacy scale

In PISA, the application of these models also permits observed student responses to test items, or more specifically the ability measures obtained from these, to be linked through regression equations to various background variables, such as gender, socioeconomic background, and so on. One outcome of these analyses is the generation of estimates of population means and other statistics that enable comparisons among groups of students between and within PISA sampled populations. And it is those comparisons among groups of students, linking performance in one or more of the cognitive test domains with various background variables, which enable policy-oriented researchers to investigate factors that may influence student performance, and to consider any implications for school management or other school organisational features, teaching and learning practices, and so on.

The design of PISA test instruments, along with the scaling and analysis of PISA data, also permits monitoring of trends in test outcomes across PISA assessment cycles. Some of the test material from PISA 2000 was also used in PISA 2003, and similarly there is test material common to the PISA 2003 and PISA 2006 assessments. This enables measures of trends to be derived, through which the monitoring of changes over time can occur. It might be expected that such changes would be negligible from one cycle to the next. However, as PISA continues across further cycles, it can be expected that adjustments to policy settings and to teaching practices in different countries might lead to more noticeable changes in PISA test outcomes over longer time periods. In the first PISA test, occurring in 2000, reading was the major test domain. Any changes in reading outcomes related to changes in policy and practice that were effected in response to PISA 2000 results (or indeed other unrelated changes) might be expected to have an impact on PISA 2009 results, when reading is next the major test domain.

Reporting of PISA outcomes

Following data collection in each PISA assessment cycle so far conducted, the OECD has produced a comprehensive report that captures the major outcomes from an international perspective. Two such reports have been produced, on the PISA 2000 outcomes (OECD 2001) and the PISA 2003 outcomes (OECD 2004) respectively.

Those main international reports have provided an overview of the PISA project, details that help understand key features of the major assessment domain (reading in the 2000 report, mathematics in 2003) and how results should be interpreted in relation to the framework, cross-country comparisons of results in the major assessment domain and their relationship with some of the key student background variables, an overview of outcomes in the minor assessment domains, an overview of the information gained from the various student-level and school-level background variables captured and differences in these among countries, international comparisons of aspects of the learning environment and the organisation of schooling, and some discussion of the policy implications of these various aspects of the study.

In reporting the literacy outcomes within the major assessment domain, emphasis has been placed on the profile of student results in each country in relation to the scales and subscales that come out of the relevant framework. In the case of mathematical literacy, results are reported for an overall mathematical literacy scale, and the results are also ‘pulled apart’ and reported for sub-scales that are based on the four ‘content’ areas of the mathematics framework. Central to the profile of student results is a set of descriptions of what students located at various points along the literacy scale would typically be able to do. The mathematics proficiency descriptions provide a clear picture of the way students are able to draw on the various mathematical competencies that are described in the framework. They describe growth in mathematical literacy in relation to an increasing student capacity to demonstrate and draw on those competencies. Further technical features of the mathematics described proficiency scales are provided in the following section. The main international OECD report places significant emphasis on the relative proportions of students in each country performing at various levels along the literacy spectrum.

In reporting on the background variables (student, family, school and system factors), differences among countries with respect to those variables are described, and the relation between some of those variables and student cognitive outcomes are analysed, in an attempt to describe how these factors play out differently in different countries.

As well as the main international reports, the OECD produces or promotes a number of additional reports. Following the first assessment cycle, a detailed report on outcomes of the assessment of reading was published (OECD 2002). A number of other reports on particular aspects of the PISA 2000 assessment outcomes have also been published by the OECD, covering such matters as student engagement, student approaches to learning, school factors related to quality and equity, and others. In addition, the OECD publishes a number of more technical documents including the technical report, a database manual, sample items, and the framework documents. A similar range of reports relating to the outcomes of the PISA 2003 survey cycle have also been published by the OECD, or are in preparation. These and other PISA publications can easily be accessed through the OECD's website at http://www.pisa.oecd.org/.

Many PISA countries also produce their own national reports, giving greater detail of various outcomes within the country, or at least providing a clearer national perspective on the results, and providing in many cases more detailed analyses and interpretations that take into account various factors operating in the particular country.

PISA's described proficiency scales

In the final sections, further technical background is provided on the approach taken to develop the proficiency descriptions that are central to the way PISA reports achievement in its different cognitive assessment domains. The methodology is outlined, and then illustrated using the example of PISA mathematics. This text borrows heavily from the chapter of the PISA 2003 technical report on described proficiency scales (OECD 2005, pp. 249-270).

The PISA test design makes it possible to use modern measurement techniques, as discussed in a previous section, 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 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 are the student ability measures used as the basis for within and between-country comparisons reported by PISA. The variables are defined in the relevant PISA literacy framework. For each of these literacy variables one or more scales are defined that 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 2 in a previous section 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.

Whereas for reporting student performance in PISA the aggregated student measures are of primary interest, for the purpose of describing growth in the scales of interest, and therefore of interpreting the meaning of the student measures found, the primary focus is on the test items and the item difficulty measures derived from the Rasch analysis.

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 was iterative – stages were revisited and the proficiency descriptions were progressively refined.

Stage 1: Identifying Possible Sub-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".

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 has been interest in considering two sub-scales, for 'scientific knowledge' and 'scientific processes', but the small number of items in PISA 2000 and PISA 2003, when science was a minor domain, meant that this was not possible. For the current survey cycle, when science is the major test domain, this matter will be revisited.

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 ongoing 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 key stakeholders will expect that the proficiency scales 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. Experts in each test domain 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.

Stage 4: Analysing Field Trial Data

For each set of scales being considered, the field trial item data were analysed using Rasch analysis to derive difficulty estimates for each achievement threshold for each item in each sub-scale.

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.).

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 that 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. The preliminary descriptions and levels were then reviewed and revised, the levels were defined, and the methodology used to associate students with those levels was applied.

Stage 7: Validating

A number of approaches to validation have been used to varying degrees. One method is 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 has been made, and further validation exercises of this kind are underway.

Defining Proficiency Levels

How should we divide the proficiency continuum up into levels that might have some utility? And having defined such 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 on the scale 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 that 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 having 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 has been developed and used for PISA.

The PISA methodology for defining proficiency levels progresses in two broad phases. The first, described in the preceding section, is based on a substantive analysis of PISA items in relation to the aspects of literacy that underpin 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 levels.

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 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 spanned by each level.

For example, students at the bottom of a level would correctly complete at least 50 per cent of tasks 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 a high proportion 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 higher level, 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. Of course 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. 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 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; 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 3 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, reference to students at both the top and the bottom of the level, and reference to items at the top and bottom of the level. Dotted lines connecting the students and items are labelled "P=?" to indicate that there is some probability associated with that student correctly responding to that item.

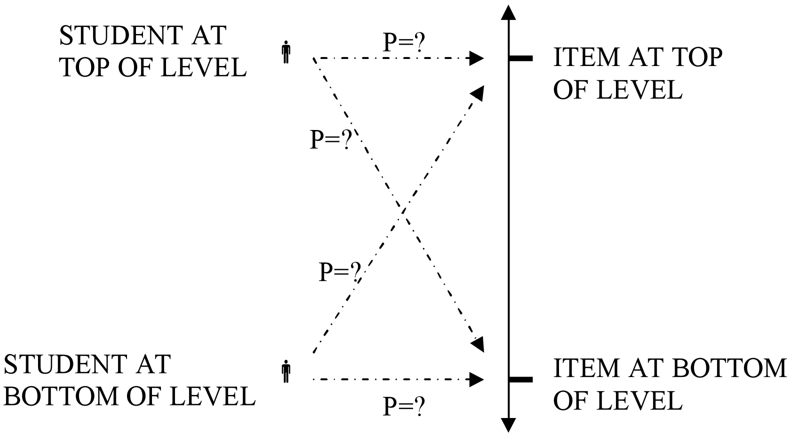

Figure 3. What it Means to 'be at a Level'

PISA has implemented the following solution: start with the preferred range of abilities for each bounded level in each scale, determined from substantive considerations (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).

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 described 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 (level 1) 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.

An illustration – described scales for PISA mathematics

Applying the processes described in the previous sections to the PISA 2003 mathematics data enabled the PISA consortium to develop described proficiency scales for PISA mathematics. This paper concludes with a presentation of some of the key results of this work.

Following data analysis and the resultant generation of difficulty estimates for all item steps, the items and item steps were associated with their difficulty estimates, with their framework classifications, and with their brief qualitative descriptions. Figure 4 shows a descriptive map of some of this information from an illustrative sample of items from the PISA 2003 test. The items referred to have all been publicly released, and can be obtained from the PISA website (http://www.pisa.oecd.org/). Each row in Figure 4 represents an individual item or item step. The selected items and item steps have been ordered according to their difficulty, with the most difficult at the top, and the least difficult at the bottom. The difficulty estimate for each item and step is given in units from the PISA scale, 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 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 4, we can see 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, we find that 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.

Code	Item name	Item difficulty on PISA scale	comments - item demands	quantity	space and shape	change and relationships	uncertainty	reproduction	connections	reflection	personal	educational/occupational	public	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			1			1		1			
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				1		1				1	
M266Q01	Carpenter Q1	687	interpret and link text and diagrams representing a real-world situation; show insight in 2-d geometrical properties; extract information from geometrical representation; calculate perimeters for compound and irregular shapes; apply routine		1				1			1		
M124Q032	Walking Q3.2	666	find a suitable strategy; multi-step problem solving; manipulation of expressions containing symbols; routine procedures; partially correct computations			1			1		1			
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				1		1			1		
M124Q01	Walking Q1	611	interpret and link picture, text and algebra; algebraic substitution; solve basic equation; single step; correct manipulation of expressions containing symbols			1		1			1			
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			1			1		1			
M413Q03	Exchange Rate Q3	586	insight into quantitative relationships; strategy: how to tackle? (problem solving); communication of conclusion and reasoning)	1						1			1	
M179Q011	Roberies 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				1		1				1	
M150Q03	Growing Up Q3	574	interpret graph in respect to rate; reasoning; communicate explanation in support of given proposition			1			1					1
M520Q02	Skateboard Q2	570	problem solving - choose a strategy; counting (combinatorics)	1				1			1			
M438Q02	Exports Q2	565	interpret graph; identify and select relevant information; link separate data and carry out routine calculation				1		1				1	
M520Q03	Skateboard Q3	554	explore possibilities to decide on which is best; interpret information; identify and select relevant information	1					1		1			
M150Q022	Growing Up Q2.2	525	link text to graphical information; locate relevant data; write conclusion correctly			1		1						1
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		1				1		1			
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	1				1			1			
M150Q01	Growing Up Q1	477	interpret graph and link to text; identify appropriate procedure; carry out simple computation (subtraction)			1		1						1
M520Q011	Skateboard Q1.1	464	interpret and link information in text and table; select and process relevant information from a table (only partially correctly)	1				1			1			
M413Q02	Exchange Rate Q2	439	interpret simple quantitative model; apply it with a simple calculation (division)				1	1					1	
M438Q01	Exports Q1	427	link representations (text and graphic); identify relevant information; read value directly from bar graph				1	1					1	
M547Q01	Staircase Q1	421	interpret simple and familiar picture; simple calculation (division by 2-digit number)		1			1				1		
M150Q021	Growing Up Q2.1	420	link text to graphical information; locate relevant data; write a partially correct conclusion			1		1						1
M413Q01	Exchange Rate Q1	406	interpret a simple quantitative model; apply it with a simple calculation involving multiplication	1				1					1	

Figure 4. A descriptive map for selected mathematics items

Near the bottom of the part of the continuum displayed here, we see 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; pp. 78-79). In those examples, students were presented with diagrams of 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, we see 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, *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, and 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, we see 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* (OECD, 2003; pp. 64-66). 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, we can characterise growth along the PISA mathematical literacy scale by referring to the ways in which mathematical competencies are associated with items located at different points along the scale.

The mathematics framework (OECD, 2003; p. 54-55) 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 is based on the definition of a number of bands or levels of proficiency in the measured literacy domain. The levels are 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. 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 to characterise typical student performance at each level. The continuum of increasing mathematical literacy 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 Figure 5.

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 5 Mathematical literacy performance band definitions on the PISA scale

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. The four sets of descriptions were then combined to produce descriptions of six levels of overall mathematical literacy, presented here in Figure 6.

Overall Mathematical Literacy	
VI	***At Level VI 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 this 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. Student 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.***
V	***At Level V 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.***
IV	***At Level IV 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.
III	***At Level III 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.***
II	***At Level II 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.***
I	***At Level I 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.***

Figure 6 Summary descriptions for six levels of overall mathematical literacy

References

OECD (2001). Knowledge and Skills for Life – First Results from PISA 2000. Paris, France: OECD.

OECD (2002). Reading for Change – Performance and Engagement across Countries. Paris, France: OECD.

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

OECD (2005). PISA 2003 Technical Report. Paris, France: OECD.

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

Routitsky, A. & Turner, R (April, 2003) *Item Format Types and their Influence on Cross-national Comparisons of Student Performance*. Presentation given to the Annual Meeting of the American Educational Research Association (AERA) in Chicago, USA.

Wu, M. L., Adams, R. J. and Wilson, M. R. (1998). *ACER ConQuest: Generalised Item Response Modelling Software*. Melbourne: ACER Press.