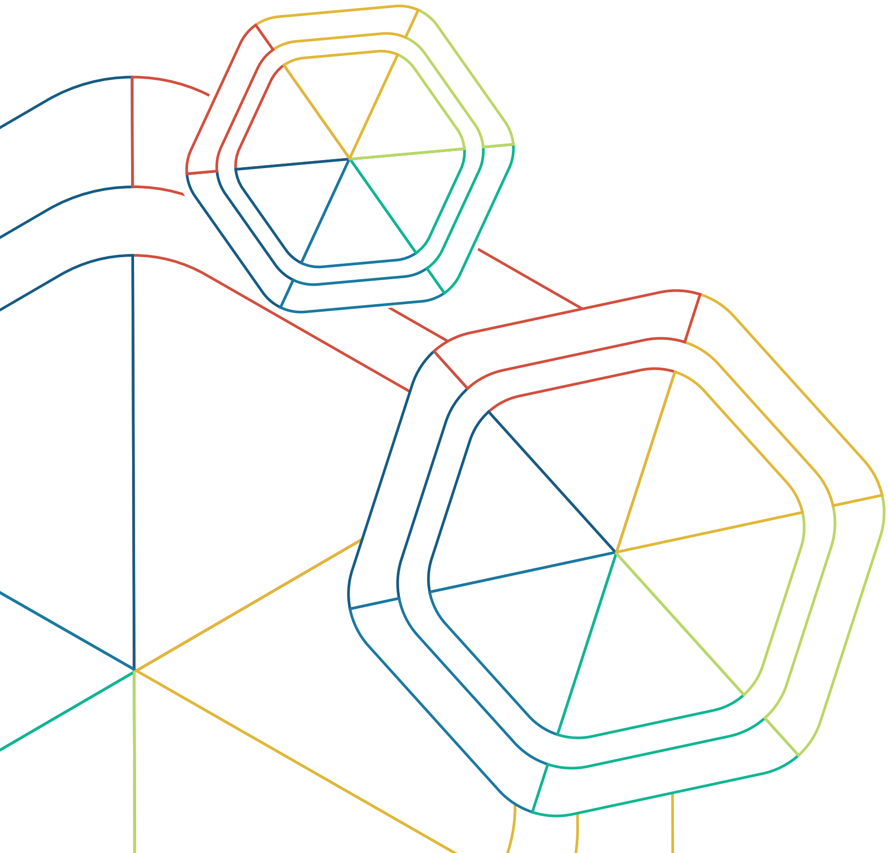
PISA for Schools

Technical Report

2022



OECD
BETTER POLICIES FOR BETTER LIVES

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Foreword

The OECD PISA-based Test for Schools (PBTS) assessment is designed for use by schools and networks of schools around the world to support research, international benchmarking and school improvement efforts. It collects information about 15-year-old students' applied knowledge and competencies in reading, mathematics and science, as well as their attitudes toward learning and school. PBTS examines how students in participating schools are prepared to meet the challenges of the future. The data collected by the assessment are an extremely valuable source of information for school principals, educators, and the wider school community.

The methodology of the PISA-based Test for Schools is complex and demanding. The PISA-based Test for Schools Technical Report describes those procedures and methodologies along with other features that enable the PISA-based Test for Schools to provide high-quality data to schools and local school administrations wanting to go further in understanding how their own individual schools perform compared with the world's leading school systems.

The first edition of this report was drafted by Noémie Le Donné, Tue Halgreen and Kelly Makowiecki. The second edition of the report was prepared by Javier Suárez-Álvarez with advice from Francesco Avvisati, Isabel Benitez, Ruochen Li, François Seyler and Tse Chi Sum, under the supervision of Tue Halgreen.

Preparation of this third edition of the report was led by Tiago Fragoso with contributions by Tanja Bastianic, François Keslair, Federico de Luca, Tomoya Okubo, Tse Chi Sum, and Nathanael Reinertsen, under the supervision of Joanne Caddy. Administrative support for the third edition was provided by Jenny Baracaldo Fernández and Stephen Flynn co-ordinated production. Acknowledgements and gratitude are also due to the volunteer reviewers from our partner organisations: Elena Govorova (2E Estudios, Evaluaciones e Investigación, Spain); Sara Ratner (Janison, Australia), Ruben Klein and Leandro Marino (Cesgranrio Foundation, Brazil).

Table of contents

Tables

Figures

Boxes

Introduction

PISA for Schools is a programme centred on the delivery and reporting of school performance, using the PISA-based Test for Schools (PBTS). Since 2019, the PBTS has been delivered as a computer-based assessment, and this third edition of the *PISA-based Test for Schools Technical Report* outlines the technical details of the administration and analysis of the computer-based version.

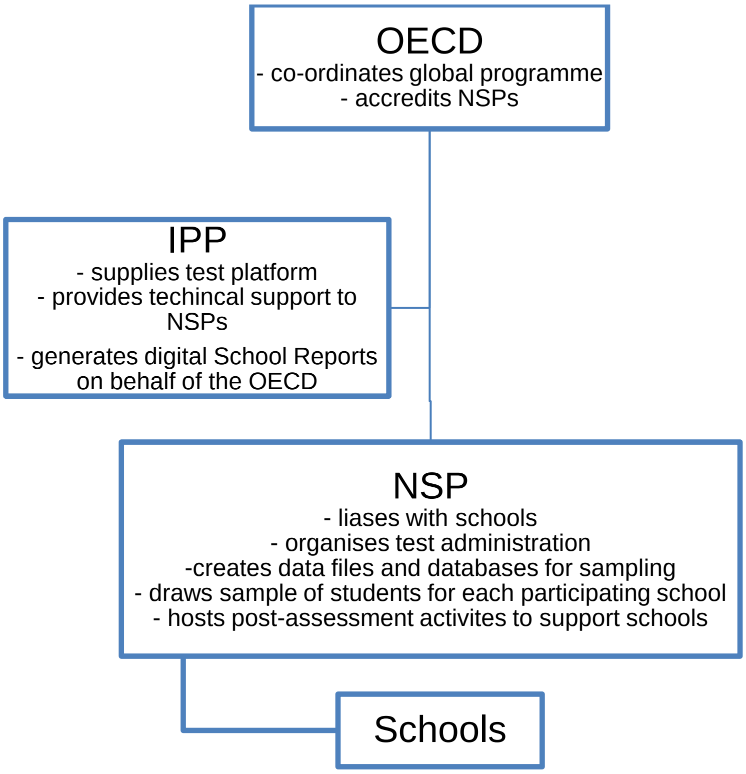
The PBTS is a student assessment tool geared for use by schools and networks of schools to support research, international benchmarking and school improvement efforts. In the United States, the assessment is known as the OECD Test for Schools (based on the Programme for International Student Assessment, PISA). The assessment tool provides descriptive information and analyses on the skills and creative application of knowledge of 15-year-old students in reading, mathematics, and science, comparable to existing PISA scales.

The assessment also provides information on how different factors within and outside school associate with student performance. Contextual questionnaires geared for students and schools are an important part of the assessment. Information about students' socio-economic backgrounds, their attitudes and interests in reading, science and mathematics, and the learning environment at school are all addressed in the assessment.

The PISA for Schools programme also provides peer-to-peer learning opportunities for educators – locally, nationally and internationally – as well as the opportunity to share good practices to help identify "what works" to improve learning and build better skills for better lives.

The PISA for Schools programme is a collaboration of several entities, co-ordinated and overseen by the OECD. The intellectual property of the test content is vested in the OECD. In each country, the test is supplied to schools by a National Service Provider (NSP) that is accredited by the OECD to do the work. The test itself is hosted on a digital platform, maintained by the International Platform Provider (IPP), currently Janison Education Solutions, Ltd. Each of the entities (OECD, NSP and IPP) have different roles and responsibilities in the delivery of the programme. These are represented diagrammatically in Figure 1.

Figure 1. Entities in the PISA for Schools programme and some of their roles


OECD
- co-ordinates global programme
- accredits NSPs
IPP
- supplies test platform
- provides techincal support to NSPs
- generates digital School Reports on behalf of the OECD
NSP
- liases with schools
- organises test administration
-creates data files and databases for sampling
- draws sample of students for each participating school
- hosts post-assessment activites to support schools
Schools

The successful operation of PISA for Schools involves significant effort from all parties, and each party acts in accordance with the needs of a variety of different stakeholders. This Technical Report has been produced in order to inform a wide variety of stakeholders about many technical aspects of the programme. Not all parts of the report will be relevant for all parties or stakeholders, but nonetheless, the report is intended to serve as a repository of knowledge so that any interested group can learn more about the PBTS.

1. Structure of this report

This Technical Report is organised into three chapters. ***Chapter 1*** provides an overview of the PBTS instruments: the cognitive test and the student questionnaire. It begins with a description of the cognitive tests' design, scope and development. It then describes the contents of the contextual questionnaire. ***Chapter 2*** describes the roles and responsibilities of NSPs, including preparations and field operations. The information in it is intended to supplement and accompany the NSP Handbook. ***Chapter 3*** is intended to provide technical insights into the analytical processes carried out by the OECD on the data collected by NSPs from participating schools. It is provided for transparency and stakeholder information, and is not intended to be a step-by-step guide to reproducing the processes.

2. Recommended further reading

There are several sources of information that are referred to in this report that readers may find it useful to consult in order to understand some topics more deeply.

- The PBTS conforms to the PISA Technical Standards, for its scores to be validly put on the PISA scale, and so the PISA Technical Standards 2018 are recommended.
- Much of the analytical methodology is informed by PISA, so the PISA 2018 Technical Report is recommended.
- It is likely that parties reading this Technical Report will be interested in conducting secondary analysis with data from PBTS. The PISA 2018 Analysis Handbook is an invaluable resource for working with PBTS Data.

1. Instrument design

Simply from its name, it is clear that the PISA-based Test for Schools is intended to be closely aligned with PISA, but separate from it. The cognitive item pool for the PBTS does not comprise PISA items. The items in the pool were developed specifically for PBTS, and linking to the PISA scale is done through incorporating a comparatively small number of PISA link items into the test forms. This chapter gives details about the design, scope and development of the PBTS items, and the equating process of the PISA scales. It also describes the student questionnaire, which is more directly derived from PISA.

1.1 Test scope, design and development

This section describes the scope and test design for the PISA-based Test for Schools (PBTS), and the processes by which the Australian Council for Educational Research (ACER) developed, over the period 2010-12, the cognitive tests for mathematics, reading and science for the paper-based version of PBTS. The OECD then linked the PBTS items administered in digital format to the PISA scale through the 2019 Linking Study (Okubo et al., 2021[1]).

a. Test scope

For each subject domain, the PISA assessment framework was used to develop the PBTS assessment. The framework defines the domain, describes the scope of the assessment, specifies the structure of the test – including items' format and their preferred distribution according to important framework variables – and outlines the possibilities for reporting results. Using the same assessment framework is vital in ensuring that the PBTS is comparable with PISA, and covers the same constructs in the same ways.

Broadly, there are three domains of literacy: mathematical, reading and science. For each domain, there are a number of cognitive processes defined. Each item in PBTS targets one domain-relevant cognitive process, and subscales are created based on the set of items associated with each process. For example, in Reading, there are three groups of cognitive processes (called *aspects* in the framework), each of which is reported as a subscale: access and retrieve; integrate and interpret; and, reflect and evaluate. Detail on the PISA assessment framework for reading, mathematics and science are published in *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy* (OECD, 2013[2]).

PBTS items are arranged in units based around a common stimulus. Many different types of stimulus are used, including passages of text, tables, graphs and diagrams, often in combination. The stimuli are classified by context and other domain-specific variables, according to the domain definitions in the assessment framework. Each unit contains from one to five items assessing students' competencies and knowledge.

A unit is identified by a short label. The units' labels consist of six characters and form the first part of the item names in the data files. The first two characters are PR, PM or PS for reading, mathematics or science, respectively. The next four characters indicate the unit within the domain. For example, PM5124 is a mathematics unit. The item names (usually nine digits) represent questions within a unit (in the current example the item names within the unit are PM5124Q01, PM5124Q02, and PM5124Q03). Thus, items within a unit have the same initial six characters plus a question number. Responses that need to be recoded into single-digit variables may have a "T" or "D" at the end of the variable name (e.g., PS7012Q07T). Anchor items from the main PISA survey are occasionally delivered in the PBTS to link the tests together and follow a similar, albeit slightly different, naming convention.

The mathematics instrument consists of 40 cognitive items (25 units), the reading assessment consists of 47 items (13 units), the science assessment consists of 54 items (25 units), for a total approximate testing time of 130 minutes.

Item formats are either selected or constructed response. Selected response items are either simple multiple choice with several responses 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 (yes/no, true/false, correct/incorrect, etc.).

Constructed response items are of two broad types: manual or expert items. Manual items require limited manual input by trained coders or, since the transition to computer-based testing, by the testing platform at the stage of processing student responses. Such items require students to construct a numeric response within very limited constraints, or only require a word or short phrase as the answer, later assigned to the predefined response categories.

Constructed response expert items require the student to propose a response and then trained expert coders interpret the student responses and assign them to one of the defined response categories. The response categories are carefully and clearly defined in the coding guide, and there is a range of possible full-credit, half-credit answers, and no-credit answers.

Table 1 shows the number of items of each type. Note that although there are a total of 141 items across Mathematics, Reading and Science, one reading item, (PR6004Q05A) is not scored separately. The response to this question is taken into consideration in the scoring of reading item PR6004Q05B.

Table 1. Item types

	Total number of items	**Number of items**				
		Constructed response expert items	**Constructed response manual items**	**Complex multiple-choice items**	**Simple multiple-choice items**	**Not scored items**
Reading	47	17	4	7	18	1
Mathematics	40	7	19	3	11	0
Science	54	20	0	16	18	0
Total	141	44	23	26	47	1

Pencils, erasers, and spare paper sheets must be provided to students undertaking the PBTS assessment. NSPs are encouraged to provide their own numbered paper sheets for drafting and calculations and to retrieve, recount and then destroy these test materials after every administration.

Calculators can be provided to students if it is the standard practice at Maths classes within the school, and the International Platform Provider's testing platform has a built-in calculator available for students to use when answering Maths or Science items.

b. Test design

The test units for mathematics, reading and science are compiled into seven item clusters; two mathematics clusters (M1 and M2), two reading clusters (R1 and R2), two science clusters (S1 and S2) and one cluster including items from all three domains (RSM). Each cluster is expected to be able to be completed by a student in 40 minutes of test time.

The cluster rotation design for the booklets are similar to designs used in PISA surveys and is shown in Table 2.

Table 2. Cluster rotation design used to form test booklets for PBTS

Booklet ID	Cluster 1	Cluster 2	Cluster 3
1	R1	RSM	M1
2	RSM	M2	S2
3	M2	M1	R2
4	M1	S2	S1
5	S2	R2	R1
6	R2	S1	RSM
7	S1	R1	M2

This is a balanced incomplete block design. Each cluster (and therefore each test item) appears in three of the three-cluster test booklets, once in each of the three possible positions within a booklet, and each pair of clusters appears in one (and only one) booklet. There are plans to expand the number of clusters and booklets in the future to include new items. When this is accomplished, there will be a larger number of clusters and booklets, but the fundamental principles of the cluster rotation design will not change.

Each sampled student is randomly assigned to one of the seven booklets administered in each school, which means each student undertakes **two hours of testing**. Booklets are randomly assigned for a number of reasons, including ensuring construct coverage at the school level and avoiding situations where two test takers in close physical proximity see the same questions at the same time.

c. Test development

Experience gained in other OECD assessments, such as PISA, has shown the importance of collaborating with an experienced test centre to help achieve conceptually rigorous material that fulfils the design requirements of the assessment framework and has the highest possible levels of cross-cultural and cross-national diversity. Accordingly, all item development was undertaken at ACER, which was responsible for the item development of the 2000 to 2012 PISA editions.

Test development for the PBTS survey commenced in 2010. Development progressed through various processes and stages, slightly different for each of the cognitive domains in which test material was required, and culminating in 2012 in the selection of items proposed for use in the main survey. This section presents the development arrangements and approaches taken by ACER to produce the testing materials.

The test development teams at ACER conducted development of items, including cognitive laboratory activities, in English. Each domain team included individuals who have been involved in the test development for the main PISA surveys.

A total of 420 cognitive items were developed by ACER in two phases between 2010 and 2012. All items were field trial tested, along with 63 PISA link items, across students from schools in Australia, Ireland, the United Kingdom and the United States. Data from the international field trial was analysed using standard item response techniques.

The results of the field trial for the 420 available items were evaluated by the expert group in terms of their substantive quality, fit to framework, range of difficulty, psychometric quality, durability and interest level for 15-year-olds.

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

- The psychometric properties of all selected items had to be satisfactory.
- There had to be an appropriate distribution of item difficulties, broad enough to generate useful measurement data at both extremes of the anticipated ability distribution of sampled students across all participating countries.

Characteristics of the item set used in the field trial, and the selected set for the main survey, are presented in the *PISA-based Test for Schools: Technical Report* produced by ACER (ACER, 2012[3]).

In selecting PISA link items, framework balance, range of difficulty, and a high level of reliability were considered as prime criteria.

d. Reporting PBTS results on PISA scales

The PISA scale for reading was developed in PISA 2000, when reading was the major domain of assessment. The scale was established so that the mean and standard deviation of the PISA 2000 scores was 500 and 100 respectively, for the equally weighted 27 OECD countries that participated in PISA 2000 that had acceptable response rates (Adams and Wu, 2002[4]).

From PISA 2003 onwards, the decision was made to report the reading scores on this previously developed scale, so the reading reporting scales used for all PISA editions are directly comparable. The value of 500, for example, has the same meaning in any of these evaluations as it did in PISA 2000.

Mathematics was the subject of major development work for PISA 2003. For mathematics, the reporting scale was determined such that the mean was 500 and standard deviation 100 for the 30 OECD countries that participated in PISA 2003. For PISA 2006, 2009, 2012 and 2015 the decision was made to report the mathematics on the PISA 2003 scale.

For science, a new scale was established in 2006. The metric for that scale was set so that the mean was 500 and standard deviation 100 for the 30 OECD countries that participated in PISA 2006. For PISA 2009, 2012 and 2015 the decision was made to report the science scores on the PISA 2006 scale.

To permit a comparison of the PBTS results with those of PISA, the decision was thus made to report:

- the PBTS reading scores on the PISA reading scale developed for PISA 2000;
- the PBTS mathematics scores on the PISA mathematics scale developed for PISA 2003; and
- the PBTS science scores on the PISA science scale developed for PISA 2006.

Further details on the various PISA reporting scales are given in Chapters 9 and 12 of the *PISA 2012 Technical Report* (OECD, 2014[5]).

Following PISA's change to computer-based testing as its main mode of administration in 2015, the PBTS changed from its previous paper-based test instruments to a digital format in the 2019-20 testing cycle. From the 2020 cycle onwards, the PBTS has been offered and delivered exclusively in a digital format.

The linking of PBTS reading, mathematics and science items in their now current computer-based forms to the existing PISA scales were performed through a Linking Study in late 2019 with four countries (Brazil, United States, Russian Federation and Spain). The methodology of PBTS was revised and improved after the 2019 testing cycle. The improved methodology applies to all participating countries from January 2020. The transformations for putting the PBTS logit scores on the PISA point scales are given in Section 3.3.b.ii.

The 2019 Linking Study had the main objective of linking the 141 PBTS items in digital form, delivered through a partner IPP platform in a computer-based assessment format, to the PISA scale. This means that PBTS computer-based results are linked to the PISA scale, and from previous links, previous paper-based PBTS results are also linked to the PISA scale but the two PBTS sets of results are not directly comparable, as mode (computer versus paper) effects were not investigated and might pose a confounding factor. Therefore, direct comparisons between paper-based and computer-based PBTS results should be avoided.

e. Digital test delivery

As reported in the previous section, since the 2020 test cycle, the PBTS has been delivered only via digital platforms. All NSPs that have joined the PISA for Schools programme since 2019 deliver the PBTS via the International Platform Provider's platform. The IPP for the period 2019-2024 is Janison Solutions Pty. Ltd.

The test content was largely unchanged between the paper-based and digital versions of the PBTS. The transition to a digital test mode was undertaken in order to enhance the security of the test materials, to enable centralisation of data collection and analysis, and to ensure the test meets the expectations of 21st-century schools.

Janison created an instance of its test platform that could be tailored to the PBTS, and continues to collaborate with the OECD in developing the platform to meet the needs of NSPs and schools participating in the PISA for Schools programme.

2. Context questionnaires

This section describes the content of the Student Questionnaire that is delivered alongside the cognitive instruments in the PISA-based Test for Schools.

Student Questionnaire

The PBTS Student Questionnaire (StQ) includes subsets of questions from four PISA Student Questionnaires: PISA 2012 (OECD, 2012[6]), 2015 (OECD, 2014[7]), 2018 (OECD, 2017[8]) and 2022 (OECD, forthcoming), and selected items from the OECD Study on Social and Emotional Skills (SSES) ([9]). PISA items are presented in the exact format and structure from their original instruments.

Questions were selected for inclusion for their comparability with recent PISA results and for their analytical relevance for school-level reporting. The first section of the questionnaire includes core questions on the student's family background and school experience and the subsequent sections explore student views and experiences learning mathematics, science and language classes with special attention to their views on reading. Further sections focus on student perception of school climate and teaching practices, and conclude with items on how students feel and behave.

Additional modules can be added upon demand from countries and changes in PISA. For instance, items from the PISA Global Crises Module (Bertling et al., 2020[10]) were added to the end of the PBTS StQ to assess how students fared during the protracted school closures caused by the COVID-19 pandemic. All additional modules and items are to be added at the end of the StQ unless agreed otherwise with the OECD.

Extra sections and national items notwithstanding, the core PBTS Student Questionnaire has the following structure:

- Section 1: about you, your family and your home
- Section 2: your view on reading
- Section 3: your <test language> learning experience
- Section 4: your mathematics learning experience
- Section 5: your science learning experience
- Section 6: your school
- Section 7: your view about your life
- Section 8: your typical behaviour

Questionnaire items are organised into units. Units consist of multiple-choice questions presenting one or several statements for each of which students are required to choose the best answer(s), or one or several constructed response questions.

A questionnaire unit is identified by a short label. All items in the PBTS Student Questionnaire are identified by a two letter identifier prefix "ST", followed by a three character code identifying its unit, followed by a code composed by Q and an item number to uniquely identify items within a unit and a two letter suffix to identify its origin when applicable.

For example, item ST197Q07HA indicates student response to their perceptions of international conflicts. Unit ST197 refers to the Global Competence unit, while Q07 means it is the seventh item in this unit in the original PISA instruments. The original PISA instrument is PISA 2018, as indicated by the HA suffix. The complete correspondence table with all units and codes can be provided by the OECD to interested National Service Providers upon request.

Three sets of questions in the PBTS Student Questionnaire are constructed response items that require the use of trained coders to interpret observed student responses and assign them to one of the defined responses. Details on how to code student responses to these occupation-related questions are provided in Section 2.3.e.

The total estimated response time to the PBTS Student Questionnaire for each student is approximately 40 minutes.

References

ACER (2012), *PISA-based Test for Schools: Technical Report*, Australian Council for Educational Research, https://www.oecd.org/pisa/aboutpisa/PISA-based%20Test%20for%20Schools%20Technical%20Report%20-%20ACER%202012.pdf. [3]

Adams, R. and M. Wu (eds.) (2002), *Programme for International Student Assessment (PISA): PISA 2000 Technical Report*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264199521-en. [4]

Bertling, J. et al. (2020), “A tool to capture learning experiences during COVID-19: The PISA Global Crises Questionnaire Module”, *OECD Education Working Papers*, No. 232, OECD Publishing, Paris, https://doi.org/10.1787/9988df4e-en. [10]

OECD (2021), *Beyond Academic Learning: First Results from the Survey of Social and Emotional Skills*, OECD Publishing, Paris, https://doi.org/10.1787/92a11084-en. [9]

OECD (2017), *STUDENT QUESTIONNAIRE FOR PISA 2018*, OECD, https://www.oecd.org/pisa/data/2018database/CY7_201710_QST_MS_STQ_NoNotes_final.pdf. [8]

OECD (2014), *PISA 2012 Technical Report, OECD, Paris*, https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf. [5]

OECD (2014), *STUDENT QUESTIONNAIRE FOR PISA 2015*, OECD, https://www.oecd.org/pisa/data/CY6_QST_MS_STQ_CBA_Final.pdf. [7]

OECD (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264190511-en. [2]

OECD (2012), *Database - PISA 2012*, http:////www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm. [6]

Okubo, T. et al. (2021), “PISA-Based Test for Schools: International Linking Study 2020”, *OECD Education Working Papers*, No. 244, OECD Publishing, Paris, https://doi.org/10.1787/ef1356ae-en. [1]

OECD (2012), *PISA 2009 Technical Report*, OECD, Paris, http://www.oecd.org/pisa/pisaproducts/50036771.pdf.

OECD (2009), *PISA 2006 Technical Report*, OECD, Paris, http://www.oecd.org/pisa/pisaproducts/42025182.pdf

2. National Operations

The objective of this chapter is to provide an overview of National Service Provider (NSP) roles and procedures in alignment to PISA-based Test for Schools (PBTS) and PISA Technical Standards, providing a technical companion to the more operational NSP Handbook, produced by the OECD jointly with the International Platform Provider (IPP).

2.1 National Service Provider specific roles and responsibilities

The National Service Provider is responsible for the implementation and testing operations of the PBTS assessment within a territory, as well as further activities derived from the assessment data, such as post-assessment workshops. The National Service Provider is accredited by the OECD following review of their technical capacity to successfully implement the PISA-based Test for Schools (PBTS).

A significant portion of the time and work invested by the NSP will be during the first year of preparations and PBTS testing operations as part of the initial validation study. The successful completion of testing operations relies on three key roles: the NSP Co-ordinator, who oversees all PBTS operations, a team of School Co-ordinators who liaise with the NSP on behalf of each school or group of schools, and the Test Administrators at the front line, administering the PBTS to students.

a. NSP Co-ordinator

The National Service Provider must identify one or more co-ordinators at their institution who will be responsible for the management of the PISA for Schools project in a given territory. Their role will be twofold: to liaise with the OECD and with the IPP in name of the NSP to oversee PBTS operations, and to co-ordinate NSP activities and PBTS administrations in their territory.

Regarding the PBTS operations, the NSP Co-ordinator:

- Liaises with the OECD and IPP on all technical and operational issues regarding PBTS administration
- Co-ordinates translation and adaptation efforts of all PBTS materials and manuals
- Compiles a list of schools to be tested and centralises communications with schools
- Manages the school data collection step, consolidating all student and infrastructure data and the production of the student samples
- Co-ordinates the IPP platform in their testing language, manages validation studies, and the scheduling and implementation of PBTS administrations
- Monitors PBTS administrations on testing days, liaising with the IPP to ensure uniform and fair testing conditions at all tested schools
- Identifies, trains and oversees School Co-ordinators, Test Administrators and Coders
- Manages marking operations

b. School Co-ordinator

The National Service Provider must identify a School Co-ordinator in each participating school. In some contexts, this person is nominated by the school principal or is a volunteer from the school staff. In others, this role could be played by a person or team of persons hired or subcontracted by the NSP to handle field operations.

School Co-ordinators co-ordinate school-related activities with the National Service Provider and the Test Administrators.

The School Co-ordinators:

- Establish the testing date and time in consultation with the NSP.
- Prepare a student list with names and relevant information of all eligible students in the school and send a version of it *without* the student names included (replacing them with an index number or unique student number) to the NSP for student sampling.
- Receive the list of sampled students on a student tracking form from the NSP and update it if necessary, including identifying transferred students, students with disabilities or limited test language proficiency who cannot take the test according to the criteria established.
- Inform school staff, students and parents if necessary about the nature of the test and the test date by sending a letter or organising a meeting, and secure parental permission if required by the school.
- Assist the Test Administrator with room arrangements for the test day.
- On test days, ensure that sampled students attend the test session(s), providing detailed attendance data to the NSP and schedule follow-up sessions if necessary.

The *School Co-ordinator's Manual* shared by the International Platform Provider describes in detail the activities and responsibilities of the School Co-ordinator.

c. Test Administrators

The Test Administrators are primarily responsible for administering the PBTS test fairly, impartially and uniformly, in accordance with international standards and PBTS procedures. To maintain objectivity, the Test Administrators are usually employed or subcontracted by the NSP, and should preferably not be school staff (OECD, 2020[11])[1].

Prior to the test date, Test Administrators need to be trained by the NSP. Training includes a thorough review of the *Test Administrator's Manual*, and the script to be followed during the administration of the test and questionnaire. Additional responsibilities include:

- Ensuring receipt of any testing material from the National Service Provider and maintaining their security and confidentiality
- Assisting the NSP or School information technology (IT) support staff for the successful installation of the digital testing platform
- Co-operating with the School Co-ordinator

[1] See PISA 2022 Technical Standard 8.1, including Note 8.1

- Contacting the School Co-ordinator prior to the test and confirming training and administration plans
- Completing final arrangements on the test day
- Conducting a follow-up session, if needed, in consultation with the School Co-ordinator
- Reviewing and updating the student tracking form
- Completing the session attendance form (a form designed to summarise session times, any occurrences during testing, etc.)
- Ensuring that all tests in their room have been successfully submitted, have been collected into their storage devices, or that all responses were successfully reconciled with the testing platform
- Ensuring the integrity of testing materials, namely that no confidentiality was breached, all testing materials are accounted for, and there is no further access to the testing platform
- Delivering or sending all test materials to the National Service Provider after the testing is carried out

The *Test Administrator's Manual* describes in detail the activities and responsibilities of the Test Administrators.

2.2 Preparing for implementation

a. Translation, adaptation and verification of the test and survey material

The PBTS assessment is intended to be available in a large number of countries with different languages, different cultures and different school systems. The PBTS follows the main PISA practice of providing the testing materials in the language of instruction in the content areas being tested.

The aim is to assess 15-year-old students' skills in three major domains: reading, mathematical and science literacy, using strictly equivalent test batteries, translated and adapted into each of the languages of instruction of the participating schools. In order to achieve this, PBTS implements strict procedures for translation/adaptation and verification of assessment materials.

The procedures firstly consist of the development of a source version of the instruments in English (see Section 1.1.c). If the language of instruction is not the source language (English), a full translation, as described here, might be needed. For both English language and non-English language applications, some degree of adaptation of the materials to local educational and cultural terminology is needed.

i. Language of instrumentation and administration

In some countries, the PBTS is to be administered in more than one language, so the procedures for translation, adaptation and validation need to be carried out for each language. The decision about choice of the test language can be made at the student, school or country level. Such situations must be discussed on a case-by-case basis and agreed by the OECD beforehand.

ii. Materials to be prepared

The key end-user materials need to be translated into the assessment language or languages so as to be linguistically equivalent to the PBTS source version. These materials comprise:

- All administered assessment items
- Coding guides for the cognitive items
- All or part of administered context questionnaires, depending on country
- School Report and Reader's Guide
- Test Administrator script
- International Platform Provider materials (NSP Co-ordinator, School Co-ordinator, Test Administrator and other manuals if applicable, as well as the user interfaces of the test platform)

iii. Procedures for preparing the materials

Prior to the validation study (see Section 2.3.b), the NSP must prepare a translation plan that clarifies the procedures to be used to develop their national version(s) and the different processes used for translator/reconciler recruitment and training.

The translation/adaptation process figures in the Post-Implementation report prepared by the NSP, which must be reviewed and approved by the OECD at the end of the testing period. Test units, questionnaire items, manuals and coding guides are initially sent to NSPs several months before the testing dates, allowing adequate time for materials to be translated (if required), adapted and verified.

A single translation of the materials should be undertaken by professional translators. The *PBTS Translation and Adaptation Guidelines*, which are based on the PISA 2018 Translation and Adaptation Guidelines (OECD, 2016[12]), contain general instructions, a number of recommendations to address common translation traps encountered when translating test materials, a list of adaptations that are desirable, acceptable or crucial when translating cognitive tests, and notes on translation and adaptation of questionnaires and manuals.

– Cognitive test material

NSPs are required to submit the translated/adapted items in units, i.e., in sets of items associated with the same stimuli. The cognitive items must be submitted along with a form documenting any proposed national adaptations for verification by the OECD.

As in main PISA, one of the most important quality control procedures implemented to ensure high-quality standards in the translated assessment items for PBTS is to have an independent team of expert verifiers confirm each national version against the English source version. The OECD establishes one verification centre to be in charge of the linguistic verification of the cognitive items.

Once the verification of the cognitive items has been approved by the OECD, the NSP incorporates the requested modifications into their materials and shares said materials through secure file transfer protocols with the IPP so booklets can be assembled for revision in the testing platform.

– Context questionnaires

As with the test material, the source versions of the context questionnaires in English (or parts thereof) are provided to NSPs for translation and adaptation into the test languages. As all the questionnaire questions come from the PISA context questionnaires, an NSP whose country has already participated in main PISA studies can use the questions translated and adapted for their country (and their language), the same rationale applying to items from the OECD Study on Social and Emotional Skills.

NSPs are permitted to add questions of national interest as national options to the questionnaires. The additional material should be placed at the end of the international modules unless otherwise agreed upon. Proposals and text for national options, as well as their placement, are to be submitted to the OECD for approval as part of the process of reviewing adaptations to the questionnaires. With the OECD's approval, NSPs can also take out questions that are not relevant to the local context, if such questions are not used in the computation of key school-level results.

NSPs are required to submit a note documenting all proposed national adaptations to questionnaire items to the OECD for approval. NSPs implement the OECD's feedback in the final version of the questionnaires, which is submitted once more in order to conduct a final check. Following feedback from the final check, NSPs make final changes to their questionnaires prior to approving test materials for testing in the platform.

- School Co-ordinator and Test Administrator manuals

The *NSP Co-ordinator's Manual*, *School Co-ordinator's Manual* and *Test Administrator's Manual* are also required to be translated into the language of instruction. English versions of each manual are provided by the IPP and translated versions are to be sent to the OECD for checking.

All survey instruments that have been translated, adapted and verified are the OECD's property and are made available to new participating countries whenever appropriate.

b. Validation study participation and outcomes

The PBTS can be made publicly available to schools once a country has successfully implemented a validation study. The purpose of the validation study is to make decisions regarding item treatment for each language and country. This means that an item may be deleted from the PBTS scaling in a country (or in a language) if it has poor psychometric characteristics in that particular country (or language). Thus, items that were empirically shown to not be appropriate for a determined national or language group will not affect student results.

The sample size for the validation study is a function of the test design and is set to achieve the standard of at least 200 student responses per cognitive item, for each language being validated. The schools selected to participate in the validation study should be as diverse as possible in terms of level of achievement, school size, intake and type. The NSP may want to consider increasing the number of students tested if it wants to deliver reports to the participating schools during the validation study, given that schools must provide a minimum of 42 valid responses in order to receive a School Report.

During the validation study, the NSP sets up a help desk. School Co-ordinators and Test Administrators are encouraged to send queries to the service so that a common adjudication process is consistently applied to all questions about test administration procedures. All practical issues that have arisen during the validation study operations, as well as the solutions proposed by the NSP to address and improve them for the publicly available study, must be documented.

The OECD analyses the validation study data, including conducting psychometric analyses of cognitive items. Analyses of cognitive items are to be performed for each language of instruction available in the country (as described in Chapter 3). Particular attention is paid to the fit of the items to the scaling model, item discrimination and item-by-country interactions (and item-by-language interactions if the test is administered in more than one language in the participating country).

2.3 Field operations

a. School and student eligibility to participate in the survey

i. Target definition

The NSP is responsible for the recruitment of schools. All schools in the country are eligible to participate in the PBTS test if they meet the minimum requirement of having 42 or more students who are between the ages of 15 years and 3 completed months to 16 years and 2 completed months at the time of assessment. The operational definition of an age population directly depends on the testing dates. A variation of up to one month in this age definition is permitted.

Recognising that in many contexts, the requirement to have 42 eligible students in a single school may be impractical, the OECD can work with the NSP to create school clusters that have a combined number of at least 42 eligible students. Schools in a school cluster must not be grouped simply by school size. They must be grouped by relevant variables they share, such as socio-economic status and geographic location. Rather than School Reports, a school cluster will receive a cluster report. This is why the schools must all share similar characteristics: if they were merely a convenient group of dissimilar schools, the report would have no practical value to the schools.

ii. Recommendations regarding scheduling of testing

Testing is not to take place unless otherwise agreed upon:

- During the first six weeks of the school year
- During the assessment period of other OECD tests such as PISA, TALIS, etc.

For logistic purposes, it is strongly recommended to offer schools to be tested during a limited period or periods (in concentrated "testing windows") rather than throughout the school year. The testing window is to be no longer than eight weeks, as specified in PISA Technical standard 1.3. (OECD, 2020[11]).

iii. Within-school exclusions

International within-school exclusion rules for students are adopted from PISA. The most recent version of the exclusion rules are given in Chapter 4 of the 2018 Technical Report (OECD, 2020[13]). Those relevant to the PBTS are specified as follows:

- *Intellectually disabled students: these students who have a documented mental or emotional disability and who, in the professional opinion of qualified staff, are cognitively delayed such that they cannot be validly assessed in the PISA testing setting. This category includes students who are emotionally or mentally unable to follow even the general instructions of the test. Students cannot be excluded solely because of poor academic performance or normal discipline problems.*
- *Functionally disabled students: these are students who are permanently physically disabled in such a way that they cannot be validly assessed in the PISA testing setting. However, functionally disabled students who can provide responses are to be included in the testing.*
- *Students with insufficient experience in the language of assessment: these are students who need to meet all of the following criteria: i) are not native speakers of the assessment language(s), ii) have limited*

proficiency in the assessment language(s), and iii) have received less than one year of instruction in the assessment language(s).

b. Sampling

Student sampling is undertaken in order to ensure the sample is representative of all eligible students in a school. The first stage in sampling is to prepare a list of all PISA-eligible students in each school that agrees to participate. The lists can be prepared at the national, regional, local or school level as data files, computer-generated listings, or by hand, depending on who has the most accurate information.

School sampling is also available and can be used to obtain PBTS results for groups of schools. For example, a school district or subnational government may wish to receive a report based on a representative sample of eligible schools in their purview. Even though this option of a "PISA for Schools Group Report" is available within the scope of the PBTS project, it will not be further detailed in this Technical Report, as it needs to be discussed with and approved by the OECD on a case-by-case basis before implementation. Whenever it is required, the same two-stage sampling procedure used in PISA is employed by an external contractor to the OECD, at additional cost to the NSP.

Since it is important that the student sample is selected from an accurate and complete list, the list must be prepared slightly in advance of the testing date. It is suggested that the list be received up to two months before the testing date so that the NSP has adequate time to select the student sample and confirm the list of students to be tested with schools.

i. Student sample size

For each school, a sample target size is set. This value is no lower than 42 students and is typically 85 students, although upon prior agreement from the OECD, schools or NSPs can use alternative values. For clarity, sampling will be explored using 85 students as the agreed upon within-school sample size.

When the school has more than 85 eligible students, a random selection of 85 students should be made. Schools that have a total fewer than 85 eligible students must test all students, unless agreed upon with the OECD. In order to achieve the minimum number of valid responses to be eligible for a School Report (42), the NSP is encouraged to test at least 55 students in each school, assuming a participation rate of 80%.

ii. Preparing a list of age-eligible students

A list of age-eligible students is to be prepared using the student tracking form that the OECD will provide. The following are all considered equally important; the list is not in order of importance:

- The list is to include all students, including those who might not be tested due to a disability or limited language proficiency. Students who cannot be tested are to be excluded from the assessment after the student sample is selected. In calculating the response rate, if no replacement for the student is forthcoming, the number of missing students is excluded from the denominator.

- The NSP is to provide a copy of the student list to schools and confirm whether all students therein will sit the test (i.e. no transferred students, students on medical leave or without parental consent if necessary, etc.).

- The student list should be as updated and recent as possible. Following PISA Technical Standards (OECD, 2015[14]), the student list should not be collected more than eight weeks in advance, unless otherwise agreed upon with the OECD.
- Students are identified by their unique student identification numbers. If there is no clear country-wide unique identification, the NSP will work with the IPP to create one.
- Student date of birth, grade and gender are to be reported on the student list.

iii. Stratification

Prior to sampling, students may be stratified. Stratification consists of classifying students into strata (or groups) according to selected variables referred to as stratification variables. Gender and Grade are stratification variables that could be used regularly in PBTS, with other potential variables to be discussed on a case-by-case basis.

Stratification is used in PBTS to improve the sampling quality, thereby making the survey estimates more reliable and ensuring all types of students are included and adequately represented in the sample. This is especially the case in contexts where there is a high level of grade repetition, causing the target students to be spread over a number of different grades.

During the sampling step, the NSP is to compute the proportions of eligible students in each stratum, thus determining the target number of students to be sampled within each group given the total sample size targeted (85 students in general). A random sample is to be taken of the corresponding number of students within each group. The recommended method for carrying out the selection is to use a ratio approach based on the expected total number of sampled students.

iv. Student replacements

If any sampled students are not capable of sitting the test (due to some special need or lack of language fluency), not available at school at the date of testing (they were transferred, on medical leave, etc.) or it is likely that the response rate is going to be below 80%, (due to parental refusals or absent students that are unlikely to attend a follow-up session) replacement students can sit the test.

The replacement students are identified as follows. For each sampled student, the student immediately preceding and following them in the stratum, which was ordered within in the stratification process, are designated as replacement students. The student immediately following the sampled student is designated as the first replacement, while the student immediately preceding the sampled student is designated as the second replacement.

Should the first student be sampled, the following two students can be selected as replacements, given that neither are part of the Main Sample. Analogously, should the last student be sampled, the preceding two can be sampled as long as neither are already present in the sample.

v. Preparing instructions for excluding students

The PBTS is a timed assessment, administered in the instructional language(s) of each participating school. However, students with limited assessment language experience or with physical, mental, or emotional disabilities should not be assessed.

The NSP will use the guidelines described in this report to develop any additional instructions that are suited to their national context in order to replace students in the main sample according to the information about students obtained from the schools.

The national operational definitions and occurrences for within-school exclusion are to be communicated to the OECD prior to testing and clearly documented in the NSP Post-Implementation Report.

vi. Sending the student tracking form to schools

As the main sample and replacements are decided, the NSP will consolidate the final student list into an instrument called the student tracking form. The School Co-ordinators will use said instrument to know which students are sampled in order to notify students, parents and teachers, to update information, and to identify students to be excluded.

The student tracking form should be sent four weeks before the beginning of the testing period. It is recommended that copies of the tracking form be kept readily accessible by the NSP and the School Co-ordinators in case the school copy is misplaced before the assessment day. In the interest of ensuring the PBTS is as inclusive as possible, student participation and reasons for exclusion and replacement are separately coded in the student tracking form.

Box 1. An example of drawing the student sample

For illustrative purposes, assume that an NSP is working under the following constraints on their PBTS administration for one of their schools:

- The school has an **eligible population** (age-eligible students at Grade 7 or above with sufficient mastery of the language(s) of testing and no impeding Special Needs) of **350 students.**
- School computer labs and NSP logistical constraints support the administration of the PBTS to up to 100 students, thus the **student sample size was fixed at 100** upon approval of the OECD.
- The school records and other information show **no significant gender or other biases** that would require stratification or oversampling.
- There is **no outstanding need for additional samples** (e.g. due to inclusion in samples designed to produce aggregated statistics for Group Reports).

Under these conditions, all students are to be selected with the same probability. Thus, the NSP can draw 100 students by a random draw (i.e. producing a **simple random sample**) or can utilise a **systematic random sample procedure**. The latter would guarantee a more even distribution of students throughout grades.

Below are the steps the NSP would follow to produce the student sample:

1. **Order the student records** by grade and then gender.
2. Calculate the **sampling interval** $I = \frac{eligible\ pop.}{sample\ size} = \frac{350}{100} = 3.5$ which is **rounded downwards** to $I = 3$
3. Draw a **random number** from the $[0,1]$ interval, RN. For instance, $RN = 0.45$
4. Sample the first selected student by selecting the $I * RN$-th student on the record, **rounded upwards**. For this example, the $RN \times I = 0.45 \times 3 = 1.5$ means that the **second student** in the records will be sampled.
5. The next selected student is the $RN \times I + I$-th. In this case, the **fifth student** in the registry will be selected.
6. The next selected student is obtained **by adding *I* to the previous number** and the procedure is repeated until all students are sampled.
7. **First replacements** are identified as the **ones immediately preceding** the selected students.
8. **Second replacements** are identified as the **ones immediately following** the selected students.

This will result in a Main Sample of 100 students, 100 first replacements and 100 second replacements.

vii. Student participation

The definition of *participant* aims to achieve two goals. Firstly, it is a final check of student eligibility designed to ensure that PBTS is assessing the same Target Population as PISA. Secondly, it ensures that the student data collected in PBTS contributes meaningfully to the school's aggregate results, especially performance and socio-economic background.

Any student who participates in the original or follow-up sessions is considered to be a participant, subject to the following conditions:

1. Student eligibility can be ascertained through date of birth and gender information;
2. Performance can be inferred through sufficient responses to the Cognitive test and;
3. There is enough information in the Student Questionnaire to estimate the student's socio-economic background. In particular, information regarding:
    a. Parental occupation and education background
    b. Student learning environment and socio-economic context at home

Given the conditions stated above, students can be considered participants as long as they reply to at least one cognitive item and there is enough information in the students' questionnaires to compute their economic, social and cultural status (ESCS). Otherwise, these students are considered non-participating and excluded from the student-level dataset and school-level statistics.

There must be enough students with valid responses to produce accurate school-level statistics, and these students must reasonably represent the diversity of students in the sample, in terms of age, gender, and grade. Thus, in order for a school to receive a School Report, it must meet the following criteria:

1. The total number of valid student responses within the school must be no less than 42.
2. No less than 80% of tested students must provide valid answers.

Schools that do not meet these requirements will not be issued a School Report unless agreed upon with the OECD. NSPs must be very explicit in their communications to the schools about the requirements regarding student participation. Informing schools about this requirement before testing and motivating students to provide as much reliable information as possible is strongly recommended.

Box 2. Examples of participating schools eligible to receive a school report

Term	Definition	Example 1:	Example 2:	Example 3:
Total number of eligible students	Number of students in the school who are eligible for testing	115	115	115
Sampled	Number of students in the school who are sampled to test	85	85	85
Valid[1]	Number of participating students who provide enough data to produce results	68	30	50
Will the school receive a School Report?		Yes	No	No
Reason		Number of valid responses is equal to 80% of students sampled	Fewer than 42 students.	Number of valid responses is lower than 80% of students sampled

Notes:

[1] The number of valid students includes those who have replied to at least one cognitive item and where student questionnaires include enough information to compute the economic and social cultural status (ESCS) of at least 80% of the students tested in each school. The number of valid students should be at least 80% of the students sampled.

Any exception to these standards requires prior approval from the OECD.

c. Packaging and shipping materials

Regardless of how materials are packaged and shipped, the following need to be sent to the Test Administrator:

- testing platform cognitive test and questionnaire access codes for the number of students sampled;
- a student tracking form;
- a session attendance form, and;
- additional tracking and attendance forms if there are multiple sealed materials for the same school (for different administrations) and to allow for follow-up sessions.

It is also recommended that the NSP issue reception and return forms to keep tallies of the materials distributed to schools and those returned.

Access codes for the testing platform are produced by the IPP for each student. It is recommended that individual access information be printed, each with a student identification number as well as the student’s name if this is an acceptable procedure within the country.

NSPs are allowed some flexibility in how the materials are packaged and distributed, depending on local circumstances, as long as strict confidentiality is maintained over all physical and electronic PBTS testing materials.

d. Receipt of materials by the National Service Provider after testing

It is recommended that the NSP establish a database of schools before testing begins in order to record the shipment of materials to and from schools, keep tallies of materials sent and returned, and monitor the progress of the materials throughout the various steps in processing administrative instruments after the testing.

It is also strongly recommended that upon receipt of materials back from schools, the counts of completed tests and unused access codes reported in the IPP platform be checked against the participation status information recorded on the student tracking form by the Test Administrator.

e. Coding of the cognitive test and of the context questionnaires

This section describes PBTS coding procedures. Overall, a substantial share of the cognitive items across reading, mathematics, and science domains requires manual coding by trained coders. It is crucial for comparability of results in a study such as the PBTS that students' responses are scored uniformly from coder to coder, and from country to country.

Comprehensive criteria for coding, including many examples of acceptable and unacceptable responses, prepared by the OECD, will be provided to NSPs in coding guides for each of the three cognitive domains: reading, mathematics, and science.

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

- Adapt or translate the coding guides as needed and submit these to the OECD for verification
- Recruit and train coders
- Locate suitable local examples of responses from test data to use in training and practice
- Liaise with the IPP to set up accounts in the marking platform for personnel working on the student questionnaire ("coder accounts") and for personnel working on the constructed response items from the cognitive instruments ("marker accounts").
- Organise the workflow of grading and marking activities and assure reliability is high through adjudication
- Mark the students' responses to the cognitive test items
- Code the students' responses to the profession items in the Student Questionnaire

f. Coding the test booklets

The coding of the PBTS items is to be carried out following a single coding with a 20% share of double marking. That means that 20% of student responses to all items are to be graded by two markers in order to assess rater reliability. Marker conflicts are to be solved through adjudication by senior markers and reliability statistics are to be notified to the OECD through the Post-Implementation Report. Traditional statistics are sufficient: percentage exact agreement (and percentage adjacent agreement where appropriate) and Pearson's correlations between markers. However, NSPs that have the capacity are encouraged to produce more sophisticated measures of agreement such as Cohen's Kappa.

The IPP’s marking platform organises coding so that all appearances of each item involved are marked together. Organising the marking this way has the substantial benefits of more accurate and consistent marking (because training and coding are more closely linked) and minimising effects of marker leniency or harshness (because markers mark across the range of participating students and schools).

i. *Staffing*

NSPs are responsible for recruiting qualified people to carry out the coding of the test booklets and the student questionnaires. Pools of experienced coders from other assessment projects can be called upon.

It is not necessary for markers to have advanced academic qualifications (such as Masters or PhDs), but they need to have professional knowledge on either mid-secondary-level mathematics and science or the language of the test, and to be familiar with ways in which secondary-level students express themselves. Teachers on leave, recently retired teachers and senior teacher trainees are all considered to be potentially suitable markers.

In the case of coders, it is not necessary for them to have advanced degrees, but they must be familiar with ways in which secondary-level students express themselves and professions that are familiar to the students and their context. College students and school staff on leave are considered to be potentially suitable coders.

ii. *Confidentiality form*

Before seeing or receiving any copies of PBTS test materials (including access to the marking platform), prospective coders are required to sign a confidentiality form, obliging them not to disclose the content of the PBTS tests beyond the groups of coders and trainers with whom they will be working. A template non-disclosure agreement (NDA) must be prepared by the NSP, who will also ensure all staff in contact with PBTS materials do so only after they have signed their respective NDAs. Individual agreements need not to be shared with the OECD, but it is strongly recommended for NSPs to store signed NDAs for a reasonable period after testing is complete.

iii. *Training*

Coders and markers are required to attend one coder training session co-ordinated by the NSP. At the training session, NSPs familiarise staff with the coding guides and their interpretation. If NSPs are not familiar with PBTS items and their coding, coding can be outsourced to an external contractor subject to approval as outlined in the NSP accreditation agreement.

It is recommended that prospective coders be informed at the beginning of training that they will be expected to apply the coding guides with a high level of consistency and that reliability checks will be made frequently by the overall supervisor as part of the coding process.

iv. *Length of coding sessions*

Coding responses to open-ended items is mentally demanding, requiring a level of concentration that cannot be maintained for long periods of time. It is therefore recommended that coders work for no more than six hours per day on actual coding, and take two or three breaks for coffee and lunch.

g. Student Questionnaire coding

Questions from several PISA Student Questionnaires and other OECD surveys were assembled to form the PBTS Student Questionnaire. The PBTS Student Questionnaire includes most of the core items from the PISA 2012 Student Questionnaire to allow for direct comparisons with PISA results and analyses from 2012 onwards. Most of these core questions have remained unchanged in subsequent editions of PISA and are part of a pool of basic questions that are retained for all PISA cycles.

Most PBTS Student Questionnaire items are marked automatically by the testing platform, but the three occupation items (ST014, ST015 and ST144) related to parental and future occupations must still be coded through the International Platform Provider's coding platform.

Those are to be coded according to the 2008 version of the International Standard Classification of Occupations (ISCO) codes following the guidelines produced by the International Labour Organization (ILO) and, more specifically, national versions usually produced by the national institute responsible for Official Statistics. The ILO guidelines do not include additional codes used in PISA and PBTS for some types of response, such as missing or vague. These additional codes are found in in the PISA 2018 codebook (OECD, 2018[15]).

If no national version is available, the NSP is to translate the ISCO code table into all languages in which they are delivering the PBTS and provide the translation(s) to the IPP so that the coding platform can be prepared in the language(s) of testing.

References

OECD (2020), *PISA 2018 Technical Report*, https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TecReport-Ch-04-Sample-Design.pdf. [13]

OECD (2020), *PISA 2022 Technical Standards*, OECD, https://www.oecd.org/pisa/pisaproducts/PISA-2022-Technical-Standards.pdf. [11]

OECD (2018), *PISA 2018 Database: Code book*, https://webfs.oecd.org/pisa2018/PISA2018_CODEBOOK.xlsx. [15]

OECD (2016), *PISA 2018 Translation and Adaptation Guidelines*, https://www.oecd.org/pisa/pisaproducts/PISA-2018-TRANSLATION-AND-ADAPTATION-GUIDELINES.pdf. [12]

OECD (2015), *PISA 2018 Technical Standards*, https://www.oecd.org/pisa/pisaproducts/PISA-2018-Technical-Standards.pdf. [14]

3. Data processing

For data processing, the National Service Provider (NSP) co-ordinates activities with the International Platform Provider (IPP), whose platform will collect students' data, and with the OECD, who is responsible for the calculation of results and generation of the final datasets. The OECD is responsible for constructing the school datasets, the context and performance variables, and for generating schools' results delivered in the School Report. Additionally, the OECD conducts a validation study in the first testing cycle of any new translation of the instrument, and has conducted linking studies to ensure the comparability of PBTS scores to the PISA scale. Creating context and performance variables requires using specialised software and psychometric methods. Results reported in the School Report are generated using the same statistical methodology employed in the production of PISA results.

This part of the report refers to methodologies employed by the OECD in the production of PBTS results, from data preparation and processing of the student responses collected by the IPP to the production of the aggregated statistics present in the School Report and other publications.

Briefly, this section aims to provide transparency for NSPs, and other stakeholders, of the methods utilised in the PBTS assessment, their rationale and alignment with PISA Technical Standards and the best methods adopted in specialised literature.

3.1 Constructing the initial PBTS database

a. Files in the initial PBTS database

The IPP will prepare the initial PBTS databases of students and schools by extracting the information from the platform. The initial PBTS database consists of three data files: two with student responses (raw cognitive data and scored questionnaire data) and one with the aggregated school statistics, with all the information necessary to create the School Report. Additional files with aggregated statistics at higher levels (e.g. subnational Regions) will also be produced as a companion to Group Reports whenever applicable.

The data file templates and codebooks will be provided by the OECD. The NSP will prepare the file according to the data file templates and codebooks, including a School Unique Identifier (SUI). The SUI is created by concatenating:

1. The three-character (Alpha-3) ISO 3166 Country Code (for example, AUS for Australia, AUT for Austria etc.[2])

[2] A comprehensive note of ISO 3166 codes can be found at: https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

2. The School Identifier (built, for example, from the School number in a district and the District number, or other national unique school identification number), and
3. The Year of Administration (4 to 6 digit code).

This SUI must be unique for each school and for each participation year, making it possible to track schools across cycles. If the NSP does not have those numbers available, the IPP will provide those unique identifiers upon request.

Special attention must be paid to schools that participate more than once, in which case the unique code should be the same except for the last two digits, which represent the year of administration.

i. Student data file

For each student who sits the PBTS, the student data file will contain the following information:

- Identification variables for the school (School Unique Identifier) and the student (Student code).
- Test booklet identification and test language.
- The student coded responses to the Cognitive items.
- The student coded responses to the Student Questionnaire.
- Student composite indexes (e.g. ESCS), performance scores (i.e. plausible values for all assessed PISA main cognitive domains and sub-domains) and weights.

The PBTS items are organised into units. Each unit consists of a stimulus (consisting of a piece of text or related texts, pictures or graphs, or other interactive material) followed by one or more questions related to said unit.

The International Platform Provider will prepare and share with the OECD two versions of the student data file for revision, the first being a raw student data file with the direct responses of the students to all cognitive items, and the second being a scored student data file with students' responses to the Cognitive test and Questionnaire, after marking and coding.

ii. School file

The School Questionnaire data file is compiled by the NSP and it must contain the following information for each school that participated in the assessment:

- Identification variables for the school (School Unique Identifier) and the student.
- The school responses on the School Questionnaire, if administered by the National Service Provider.
- The testing date.
- School record and attendance information such as number of eligible students, number of sampled students and number of tested students.

The National Service Provider will send the school data file up to one month after administration has been concluded so work can resume on the generation of the School Report and PBTS data products. In the case that part of the abovementioned school data is collected through the IPP's platform, the NSP will liaise with the IPP to prepare the data extract to be utilised by the OECD.

b. Records in the database

i. Records included in the database

The student raw response file shared with the NSP will contain records of all eligible PBTS students who attended test sessions and replied to at least one cognitive item. Eligible students who attended the questionnaire session are included in the Student Questionnaire file if they provided at least one response to the Student Questionnaire, their sex is known, and the father's or the mother's occupation is known from the Student Questionnaire. Students contained in both files are deemed valid responses and counted towards a school's eligibility to obtain a School Report.

The school file should contain records of all participating schools with 42 or more valid student responses collected in the assessment sessions adding up to no less than 80% of the number of originally sampled students.

ii. Records excluded from the database

Students who do not reply to any of the cognitive items (i.e. all of whose responses to the cognitive items are missing) will be dropped from the scoring process.

The following records should also be excluded from the student file:

- Sampled students who are later reported as not eligible, students who are no longer at school, students who are excluded for physical, mental or linguistic reasons, and students who are absent on the testing day.
- Students who refused to participate in the assessment sessions.

c. Representing missing data

The coding of the data must distinguish between four different types of missing data:

- Item level non-response: 9 for a one-digit variable, 99 for a two-digit variable, 999 for a three-digit variable, and so on. Missing codes are shown in the codebooks. This missing code is used if the student or school principal is expected to answer a question, but no response is actually provided.
- Not-administered: 7 for a one-digit variable, 97 for a two-digit variables, 997 for a three-digit variable, and so on. Generally, this code is used for cognitive and questionnaire items that are not administered to the students and for items that are deleted after assessment because of platform or translation errors.
- Not-reached items: all consecutive missing values clustered at the end of test session are replaced by the non-reached code, '6', except for the first value of the missing series, which is coded as item level non-response.

d. Merging the data files

Once the data files are well-prepared and organised, the student scored responses file must be merged with the student and school questionnaires data files using the student and School Unique Identifier variables. The following procedures are performed by the OECD in data cleaning:

- Perform quality assurance and check the consistency of the automatic grading algorithms for PBTS items for the current administration.

- Resolving cases of unmatched students or schools in the data files.

3.2 Procedures for scaling cognitive data

Procedures for scaling cognitive data must be used in two cases: When assessing the psychometric properties of items administered during the validation study and when constructing student performance variables for reporting purposes.

In both cases, the OECD team uses the same Item Response Theory (IRT) models used in PISA 2015 onwards; the two parameter logistic model (2PL) for dichotomous items and the generalised partial credit model (GPCM) for items with more than two response categories.

This section first provides, in technical terms, an overview of the methodological background of these models, describing the specific model used for analysing items from the field trial and the one used for computing students' performance scores.

a. Psychometric models used in PBTS

As for main PISA surveys, the generalised partial credit model as described in the 2018 PISA Technical Report (OECD, 2020[13]) must be used to scale the PBTS cognitive data. This model results from the combination of an item response model and a population model. This section presents in rather technical terms the features of the general model and describes the two specific forms used for item analyses and for the construction of performance variables.

i. Pre-scaling checks

Before any Item Response Theory (IRT) model parameter scaling takes place, preliminary checks of response behaviour are performed using Classical Test Theory (CTT) methods. This classical approach, in opposition to the "modern test theory" provided by IRT models, investigates student behaviour in less definition, by examining their aggregated response throughout the test (i.e. using their sum scores) and throughout items (e.g. through percent-correct statistics and point-biserial correlations).

This first CTT analysis is performed on every item that is part of the PBTS, regardless of their precedence, and can indicate aberrant behaviours that arise from inadequate translation or adaptation of test materials, inadequate grading of open-ended test responses, or issues in the collection of response data through the testing platform.

Percent-correct statistics, or pass rates, provide a first estimate of item difficulty and provide a first measure of quality assurance of the instruments and marking process. Indeed, if an item presents a low pass rate, fewer students managed to answer the item correctly (or equivalently, to achieve a partial or full credit score), meaning that the item might be a difficult one.

The previous conclusion relies on the assumption that the item accurately measures student proficiency and has been faithfully translated into the testing language and authored into the test platform. Pass rates provide some preliminary evidence in investigating this assumption: if an item known to be easy (e.g. due to expert review or in light of previous PBTS administrations) but had low pass rates in the current administration, the behaviour could be attributable to translation or test platform issues. Similarly, low rates in partial credit categories relative to high rates in the full credit category might be indicative of marking that is not taking into account all the nuances indicated in the coding guides.

Along with item pass rates, point-biserial correlations are also considered for every item response category. Point-biserial correlations are a measure of association between a given item response and total student score. Correlations are calculated for every response category, and the expected behaviour is that high and positive correlations are observed between scores and correct alternatives, while low and negative values are observed for partial or incorrect response categories.

Deviations from this expected behaviour can also be used to flag potential translation and platform issues. Indeed, if a correct alternative presents low (or negative) point-biserial correlations in a given sample, it might be indicative of translation issues (since high performers might not be managing to understand what the item asks of them), platform issues or psychometric issues linked to item fit. As a general guide, the point-biserial correlation for the key in a four-option multiple choice item should ideally be 0.2 or above.

Once test data are thoroughly checked in the aggregate through CTT, the model assumption of within-country alignment of the constructs is then verified through Factor Analysis models. These models assume a single latent trait (identified with whenever the PBTS is being linked to the PISA scale3 through student proficiency in PISA and in the PBTS) influences student response through a categorical factor analytic model that takes general (i.e. international) and group-specific (i.e. national or language-based) parameters into account.

The main statistical hypothesis investigated with this modelling is whether the latent scale is invariant throughout the assessed groups, meaning that proficiency is uniformly measured, notwithstanding corrections to account for group-specific behaviour. The method of choice for PBTS data is the alignment method (Asparouhov and Muthén, 2014[16]) that allows for the possibility of partial measurement invariance through multi-group modelling. This partial invariance approach provides a first measure of overall (i.e. jointly considering all assessed items) scale invariance across national groups for the PBTS during its Linking Studies.

Given adequate psychometric behaviour in the aggregate, both in item behaviour and scale invariance, student data will be scaled to IRT models, which are the theme of the following section.

ii. Item Response Theory (IRT) models

As the PBTS tests contain both dichotomous items (having two possible scores) and polytomous items (having more than two possible scores), the item response model used is a generalised form of the 2-parameter logistic Item Response Theory (IRT) model. In the interest of precision (and brevity), the model used for dichotomous items will be referred to as two parameter model (2PL) whereas the model used for polytomous items will be referred to as Generalised Partial Credit Model (GPCM).

In the GPCM the probability that student j with ability θ will obtain a score of r on item i and measured as X_{ij} is expressed as follows when used for cognitive items in PBTS

$$P_{ijr}(\theta) = P\left(X_{ij} = r \middle| \theta, a_i, b_i, d_{i0}, \dots, d_{iM_i-1}\right) = \frac{exp\sum_{k=0}^{r}(Da_i(\theta - b_i + d_{ik}))}{\sum_{k'=0}^{M_i-1} exp\sum_{k=0}^{k'}(Da_i(\theta - b_i + d_{ik}))},$$

provided that i has M_i steps, and r is the number of steps successfully completed by the students or the number of credits obtained (e.g. partial or full credit). An equivalent but different expression is used for the questionnaire items (see Box 3).

[3] Through its Linking Studies to keep the quality of the correspondence between PBTS and PISA results.

The M_i possible response values range from 0 (no steps completed or no credit) to $M_i - 1$ (all steps completed or full credit obtained), and the steps are ordered; that is, for a given item, a higher score (a higher value of r) reflects higher ability. θ denotes the person's latent trait, the item parameter b_i gives the location of the item on the latent continuum or difficulty of the item and d_{ir} are called step difficulties: d_{ir} is the difficulty of step r of item i. D is the scale constant and is equal to 1.7.

The probability of obtaining a particular score on a particular item can be generalised to the probability of a response pattern to all items of the domain. For each item $i = 1, \ldots, I,\ \ x_j$ are collated together into a single matrix X, called the overall response pattern.

Suppose the response vector contains $x_j = [2,1,1,2,0,1]$, then the binary response pattern $\boldsymbol{u}_j$is described as

$$\boldsymbol{u}_j = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

In the parameter estimation procedure, the binary response data $\boldsymbol{u}_j$ is used instead of $\boldsymbol{x}_j$.

The probability that a student with ability θ will obtain a particular response pattern $\{x_j\} = (x_1, \ldots, x_I)$ is thus modelled as

$$f(x_j|\theta, \Delta) = P(\{X_j = x_j\}|\theta, \Delta) = \prod_{i=1}^{I} \prod_{r=0}^{M_i-1} P_{ijr}(\theta)^{u_{ijr}},$$

where $\boldsymbol{\Delta} = (a_1, \ldots, a_I, b_1, \ldots, b_I, d_{11}, \ldots d_{IM_i-1})$ represent all item parameters (discriminations, difficulties and step parameters) and the overall model likelihood can be finally obtained by the product of all observed response patterns.

iii. The conditional model

This section explains the conditional (or population) model: a multilevel Item Response Theory (IRT) model in which student covariates are incorporated into the model for student cognitive performance, resulting in the generation (or multiple imputation) of plausible values.

Following PISA methodology (OECD, 2020[13]), the model is estimated in three steps:

1. *Item calibration* (IRT scaling): The responses, consisting of dichotomously and polytomously scored items, are used to estimate the item parameters for all non-fixed item parameters from the test.
2. *Population modelling using latent regression and plausible value imputation*: At this stage, the item parameters are assumed known and fixed at their estimated values obtained at Step 1, and then a matrix of regression coefficients (Γ) and error covariance matrix (Σ) is estimated based on the cognitive responses and the student questionnaires data. With these values, five plausible values per cognitive domain per student are generated. The Plausible Values represent the latent student proficiency distribution and they are used for the production of all aggregated performance statistics.
3. *Variance estimation:* in order to obtain variance estimates for school means and other statistics, a replication approach is employed to estimate sampling as well as imputation variability.

A large number of student background variables are collected in PBTS, so introducing all as covariates in latent regression would result in a large and potentially unstable model due to multicollinearity. Thus, only some key variables are directly included into the regression model, whereas the bulk of student variability is included through summaries obtained via Principal Component Analysis.

The Principal Components (PCs) are obtained by dummy coding Student Questionnaire responses and selecting the number of PCs that explain 80% of variability and each of the PCs is orthogonal to the rest, as to keep models manageable while still providing information about response variability and non-response patterns.

Thus, the latent regression model used in PBTS uses the following covariates in all countries:

- Indicator variables for Booklet, Grade and Gender
- Highest parental occupation index (HISEI)
- School size variables
- Not-reached items ratio
- Principal Components

which will be collectively referred to as the covariates matrix Y with covariate vectors y_j, for the $j = 1, \dots, N$ students.

The latent regression model is parametrised by assuming a multivariate normal distribution as the distribution of latent traits. Namely, for the S latent traits of interest (for which it is assumed $S = 3$ for the imputation of main cognitive domains and $S = 5$ for the sub-domains), the regression level is specified by

$$\boldsymbol{\theta}_j = (\theta_{j1}, \dots, \theta_{jS}) \sim N_S(\boldsymbol{\Gamma}\boldsymbol{y}_j, \boldsymbol{\Sigma}),$$

for the regression parameters matrix $\boldsymbol{\Gamma}$, covariates matrix Σ defined above and N_S denotes a multivariate normal distribution.

Plausible values are then drawn from the posterior probability distribution for the proficiency of student $j = 1, \dots, N$

$$P(\boldsymbol{\theta}_j | x_j, y_j, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$$

where x_j and y_j are the cognitive response and covariates vector for student j respectively.

The connection between posterior density function and the IRT model can be understood applying conditional probability rules leading to

$$P(\boldsymbol{\theta}_j | \boldsymbol{x}_j, \boldsymbol{y}_j, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto P(\boldsymbol{x}_j | \boldsymbol{\theta}_j, \boldsymbol{y}_j, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) P(\boldsymbol{\theta}_j | \boldsymbol{y}_j, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) = P(\boldsymbol{x}_j | \boldsymbol{\theta}_j) P(\boldsymbol{\theta}_j | \boldsymbol{y}_j, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$$

from which the first factor in the multiplication on the right hand side is the cognitive items' model likelihood function, obtained by the product of the item response likelihoods whereas the second is the multivariate normal probability density assumed for θ_j.

The procedures stated above are in close alignment with the ones utilised to generate PISA results. The interested reader is directed to the Chapter 9 of the PISA 2015 Technical Report (OECD, 2017[17]) for further details and references.

b. The validation study and score generation steps

This section will explore two further technical tasks performed by the OECD to produce PBTS results. The first regards the technical aspects of the validation study, used to ascertain the adequacy of the PBTS instrument in a new administration. Non-technical aspects of the validation study have been described in Section 2.2.b. *Validation study participation and outcomes*. The second task is the equating of scores to the PISA scale, described in Section 1.1.d *Reporting PBTS results on PISA scales*.

i. Item analyses of the Validation Study (field trial) data

National item analyses of the field trial data are performed separately, country by country, and within country, language by language if applicable, using unweighted data.

For the item analyses, the cognitive item responses are used to examine the fit of the IRT model. In most cases, it will be assumed that students have been sampled from a normal distribution with constant but unknown mean and unknown variance. In the item response level of the model, item parameters will be freely estimated and not anchored at their international values if the model-data fit of the item is poor.

The outcomes of the item analyses are used to make a decision about how to treat each item in the participating country and for each language. This means that an item may be deleted from the scaling in a particular country and in particular languages if it has poor psychometric characteristics in this particular country and those particular languages. Conversely, if there is adequate psychometric fit but there is also differential response behaviour in said language or country, a specific set of item parameters could be adopted to use as much information from student responses as possible.

When reviewing the national item analyses, particular attention is paid to the fit of the items to the scaling model, item discrimination and item-by-country or item-by-language interactions. Three types of item analyses are performed:

1. Psychometric characteristics of the PBTS items in the pilot instruments;
2. Item response model fit
3. Differential Item Functioning (DIF): by country (international vs. country parameters)

The outcomes of these analyses, a national list of DIF items and decisions regarding item treatment are made by the OECD during the analysis of the validation study.

- Differential item functioning (DIF)

DIF analysis is central to the process of psychometric validation of tests and questionnaires. International guidelines on educational measurement and test development demand that DIF analysis be carried out to ensure construct equivalence (International Test Commission, 2013[18]).

In the same way as PISA 2018, the approach for identifying DIF in PBTS is based on the root mean square deviation (RMSD) fit statistics. This measure quantifies the magnitude and direction of deviations in the observed data from the estimated item characteristic curves for each item. A cut-off value of 0.12 is used to signal an item with DIF.

ii. National list of DIF items

For each language, the OECD will compile all the items identified as DIF. They are asked to check them carefully for any translation or printing errors. After verification of the DIF items, upon consultation with the National Service Provider, the OECD team will make decisions regarding their treatment for the student score generation for the field trial data and for future administration of the survey. It is recommended that items with translation and display problems be discarded for the student score generation from the field trial data and improved for future administration of the survey.

iii. Student score generation

A School Report presents a school's results from the PBTS and compares the students' performances in three subjects (reading, mathematics and science) with performances of peers in countries and economies that took part in past PISA surveys. For reporting purposes, more than 50 performance variables must be constructed and included in the database: five plausible values of student's performance score and five plausible levels of student's proficiency for each of the three domains of assessment, as well as each of the three sub-domains of each domain.

- Omitted responses

The PISA-based Test for Schools will treat the number of missing responses with a differentiation between item-level non-responses and not-reached responses. A response is coded as an item-level non-response if the student was expected to answer a question but provided no response. All consecutive missing values clustered at the end of a test session are replaced by the not-reached code, except for the first value of the missing series, which is coded as an item-level non-response.

Therefore, for the single missing value at the end of the test session and for the first missing value of consecutive missing values (i.e., >= 2 missing values) at the end of the test session, the missing values are coded as missing. All non-first missing values in a missing series at the end of the test session are coded as not-reached. The number of not-reached items is used as a source of background information in the generation of plausible values.

- Generation of the plausible values

The PBTS conditioning variables (vector *Z* in Equation 2) are prepared using procedures based on those used in PISA. All available student-level information, other than their responses to the items in the booklets, is used either as direct or indirect regressors in the population model. The preparation of the variables for the conditioning proceeds as follows. Variables for booklet identifier are represented by deviation contrast codes and are used as direct regressors. Each booklet is represented by one variable, except for Reference Booklet 7. Booklet 7 was chosen as the reference booklet because it includes items from every domain. The difference between the simple contrast codes that were used in PISA 2000 and PISA 2003 was that with deviation contrast coding, the sum of each column is zero, whereas for simple contrast coding, the sum is one. The contrast coding scheme is given in Annex B of PISA's Technical Reports. Further information can be found in PISA 2012 Technical Report (OECD, 2014, p. 157[19]). Using this method, the imputation of abilities for students who did not respond to any science or reading items is based on information from all booklets that have items in a domain, and not simply from the reference booklet, as in simple contrast coding. Other direct variables in the regression are gender (and missing gender, if any entries are missing), grade, parents' highest occupational status (HISEI) (see Section b), school dummy variables (with the largest school as a reference; '-1' in all dummies), and the ratio of not-reached items.

All other categorical variables from the Student Questionnaire are dummy coded. All dummy variables, the numerical variables age of school entry (09-ST06), age of arrival in the country (15-ST21), as well as the recoded numerical variable (AGE) are analysed in a principal component analysis. The number of component vectors that must be extracted and used in the scaling model as regressors is country-specific and must explain 80% of the total variance in all the original variables. Therefore, as many components as are needed to explain 80% of the variance must be extracted. This indicates that the number of components can be different in each country, depending on the influence of the regressors on the estimation of the plausible values. Whenever possible, standardised regressors should be introduced to make the interpretation of the outcomes easier.

- Transforming the plausible values to PISA scales

For PISA surveys, the reading, mathematics and science results are each reported on the scales that were established when the respective domain was a major domain. For reading, the reference scale was established for PISA 2000, for math it was for PISA 2003, for science it was for PISA 2006. So, for instance, for PISA 2012, the reading results are thus reported on the PISA 2000 scale, the mathematics results on the PISA 2003 scale, the science results on the PISA 2006 scale. This was made possible because the new PISA tests were equated to the former PISA tests. In order to facilitate the interpretation of scores assigned to students, the PISA reading, mathematics and science reference scales were designed to have an average score of 500 points and a standard deviation of 100 across OECD countries. Transformation coefficients for the transformation of plausible values to the (500, 100) PISA scale are provided in the Chapter 12 of the PISA 2015 Technical Report (OECD, 2017[17]).

- Generation of the proficiency levels

In order to render PISA results more accessible to educators, proficiency scales have been developed for the assessment domains. Since these scales are divided according to levels of difficulty and performance, both a ranking of student performance and a description of the skill associated with that proficiency level can be obtained. Each successive level is associated with tasks of increased difficulty. In PBTS, six levels of proficiency, as defined in former PISA surveys, are used for each domain of assessment. The cut-off points that frame the proficiency levels in reading, mathematics and science are presented in Table 3.

Table 3. Lower score limits for the proficiency levels in reading, mathematics and science

Domain	Lower score limit					
	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Reading	334.75	407.47	480.18	552.89	625.61	698.32
Mathematics	357.77	420.07	482.38	544.68	606.99	669.30
Science	334.94	409.54	484.14	558.73	633.33	707.93

Source: OECD (2014), *PISA 2012 Technical Report*, Table 15.1, p. 297; OECD (2012), *PISA 2009 Technical Report*, Table 15.1, p. 266; OECD (2009), *PISA 2006 Technical Report*, Table 15.1, p. 293.

Proficiency levels must be derived from the plausible values. Five plausible proficiency levels must be assigned to each student respectively according to their five plausible values.

3.3 Procedures for constructing contextual variables

a. Overview

The PBTS Student Questionnaire includes numerous items on student characteristics, student family background, and student perceptions. Responses to the items are transformed to be displayed in the School Report. Some of the items are designed to be used in analyses as single items (for example, gender). However, most questionnaire items are designed to be combined in some way in order to measure latent constructs that cannot be observed directly. To these items, transformations or scaling procedures must be applied to construct meaningful indices.

This section describes how indices, used in the analyses of the School Report, must be constructed. Three different kinds of indices can be distinguished:

- Simple indices: these indices must be constructed through the arithmetical transformation or recoding of one or more items.
- Scale indices: these indices are constructed through the scaling of several items. Typically, scale scores for these indices are estimates of latent traits derived through Item Response Theory (IRT) scaling of dichotomous or polytomous items.
- Complex indices: these indices are derived from several other indices.

All indices have been used in main PISA surveys and must be constructed following the same methodology.

b. Simple questionnaire variables

i. Grade

The relative grade index (GRADE) is computed to capture between-school variation. It indicates whether students are in the school's modal grade f (value of 0) or whether they are below or above the modal grade (+x grades, -x grades). The information about the students' grade level was taken from the Student Questionnaire (ST001), whereas the modal grade was defined by the country and documented in the student tracking form.

ii. Student age

The age of a student (AGE) is calculated as the difference between the year and month of the testing and the year and month of a student's birth. Data on students' age were obtained from both the questionnaire (ST003) and the student tracking forms. If the month of testing was not known for a particular student, the median month for that country was used in the calculation. The formula for computing AGE was

$$\text{AGE} = (100 + \text{Ty} - \text{Sy}\ ) + (\text{Tm} - \text{Sm}\ )/12$$

where Ty and Sy are the year of the test and the year of the student's birth, respectively, in two-digit format (for example 06 or 92), and Tm and Sm are the month of the test and month of the student's birth, respectively. The result is rounded to two decimal places.

iii. Grade repetition

The grade repetition variable (REPEAT) is computed by recoding variables ST127Q01TA, ST127Q02TA and ST127Q03TA. REPEAT took the value of 1 if the student had repeated a grade in at least one International Standard Classification of Education (ISCED) level, and the value of 0 if "No, never" was chosen at least once, given that none of the repeated grade categories were chosen. The index is assigned a missing value if the student did not tick any of the three categories in any levels.

c. Questionnaire scale indices

Some of the scales that have been implemented in the PISA 2015 questionnaires, such as science self-efficacy and instrumental motivation in science, can be linked to the respective scales administered in PISA 2006, via a common calibration-linking procedure. For this purpose, international item and person parameters were obtained from a Generalised Partial Credit Model (see Box 3 for further explanation) in a single analysis based on data from all students in all countries from both cycles (2006 and 2015) using the mdltm software (von Davier, 2008[20]). For each scale, only students with a minimum number of three valid responses were included. Students were weighted using the final student weight, and each country in each cycle contributed equally to the estimation. Additional analyses on the invariance of item parameters across countries, languages and cycles were conducted and unique parameters were assigned if necessary. Once this process was completed, weighted likelihood estimates (WLEs) for all examinees were obtained and the OECD mean and standard deviation of the newly constructed WLEs were matched to the OECD mean and standard deviation of the original 2006 WLEs, by applying the linear transformation equation of the form

$$WLE^{*}_{2015} = A * WLE_{2015} + B$$

in which

$$A = \frac{SD_{2006,original}}{SD_{2006,new}}$$

$$B = M_{2006,original} - A \times \ M_{2006,new} \, .$$

This procedure links the 2015 data for each of the trend scales to the respective scale established in PISA 2006. The correlations between the original and new WLEs for PISA 2006 indicated that all scales could be satisfactorily recovered. This is particularly encouraging, since the scaling model changed from the Partial Credit Model in previous cycles of PISA to the Generalised Partial Credit Model in 2015. Further description of these analyses can be found in Chapter 16 of the PISA 2015 Technical Report (OECD, 2017[17]).

PISA 2006 parameters estimated with the Partial Credit Model are therefore used in PBTS for science self-efficacy and instrumental motivation in science parameters. The parameters of the other questionnaire scale indices will also remain the same as in previous PBTS cycles, obtained from the original PISA cycle (i.e. PISA 2012 parameters for instrumental motivation in mathematics). The only Student Questionnaire indicator that has new parameters is the ESCS (see next section for further explanation). In sum, the scaling methodology in PBTS remains the same as for trend comparisons in PISA, making the analysis consistent between different cycles and comparable with PISA 2015.

- *Scaling methodology*

Questionnaire scale indices are derived from student responses to several items, using a scaling methodology similar to that used for generating students' performance scores. In the scaling model, θ no longer represents student's ability in a domain but a student's latent trait. The multinomial logit model (without conditioning) must be fitted to student response data. When using the unconditional item response model to estimate students' latent traits, item parameters ξ must be anchored at their international values. In the following subsections, international item parameters (obtained from PISA calibration samples) are provided for each index.

For each student, it is then possible to specify a posterior distribution for the latent trait, given by:

$$h_{\theta}(\theta; \xi, \mu, \sigma^2 | X) = \frac{f_X(X; \xi | \theta) f_{\theta}(\theta | \mu, \sigma^2)}{f_X(X; \xi, \mu, \sigma^2)}$$

This posterior distribution of student latent trait must be used to generate weighted likelihood estimates (WLEs) of each student's latent trait (and not plausible values as for students' performance scores).

WLEs must finally be transformed to an international metric with an OECD average of zero and an OECD standard deviation of one. The transformation must be achieved by applying the formula:

$$\theta' = \frac{\theta - \bar{\theta}_{\mathrm{OECD}}}{\sigma(\theta_{\mathrm{OECD}})}$$

where θ' is the student's trait score in the international metric, θ the original WLE in logits, and $\bar{\theta}_{\mathrm{OECD}}$ is the OECD mean of logit scores with equally weighted country samples. $\sigma(\theta_{\mathrm{OECD}})$ is the OECD standard deviation of the original WLEs.

d. Indices included in datasets, but not included in School Reports

In the past, School Reports from the PBTS included a number of indices. After feedback from school principals, it was found that these indices were difficult to interpret and thus not useful to include in the School Reports. Nonetheless, these indices are used in the principal component analysis (PCA) that identifies regressors to be used in the prior distribution for the conditional (i.e. population) model used for analysing the cognitive items. Because of their fundamental importance to the analysis, they are still included in the datasets provided to NSPs, even though they are not provided directly to schools. These indices will be of great utility to analysts looking to conduct secondary analysis, and so they are described in this section.

i. Mathematics self-efficacy

Eight items are used in the PBTS as well as in PISA 2012 to measure mathematics self-efficacy (MATHEFF). The four response categories are "Very confident", "Confident", "Not very confident" and "Not at all confident". All items must be reversed, so the higher difficulty corresponds to the higher level of confidence. For this index, item difficulties ranged from a comparatively easy one, "Solving an equation like 3x+5= 17", to more difficult ones, such as "Finding the actual distance between two places on a map with a 1:10 000 scale" and "Calculating the petrol consumption rate of a car". The scaling procedure of this index thus takes into account the fact that students feel more confident in solving linear equations than they feel applying rates and proportions to real life situations.

ii. Instrumental motivation in mathematics

Four items are used in PBTS as well as in PISA 2012 and PISA 2003 to measure instrumental motivation for mathematics (INSTMOT). The response categories vary from "Strongly agree", "Agree", "Disagree", to "Strongly disagree". All items must be reversed, so the higher difficulty corresponds to the higher level of motivation. For this index, item difficulties do not vary considerably.

iii. Science self-efficacy

Eight items measuring students' science self-efficacy (their confidence in performing science-related tasks) are included in PISA 2006 and remain in PISA 2015. These items cover important themes identified in the science literacy framework: identifying scientific questions, explaining phenomena scientifically and using scientific evidence. All items must be reverse coded for IRT scaling so that positive WLE scores on this new index indicate higher levels of self-efficacy in science.

iv. Instrumental motivation in science

Five items measuring the construct of instrumental motivation are included from PISA 2006 and used to study the trends between PISA 2006 and PISA 2015. All items are inverted for IRT scaling: positive WLE scores on this new index indicate higher levels of instrumental motivation to learn science. In PISA 2015, there are 4 items instead of 5 (item ST35Q03 was dropped in PISA 2015). However, the transformation equation links the 5 items from PISA 2006 to the 4 items from PISA 2015. Therefore, instrumental motivation in science can be computed with 5 items, using the parameters from PISA 2006.

v. Disciplinary climate in language of instruction

This scale provides information on disciplinary climate in the classroom (DISCLIMA) included from PISA 2009. There are five items in this scale, each with four response categories varying from “Strongly disagree”, “Disagree”, “Agree” to “Strongly agree”. The items in this scale must be reverse coded (i.e., higher WLE’s on this scale indicate a better disciplinary climate and lower WLE’s a poorer disciplinary climate). Similarly, positive item difficulties indicate aspects of disciplinary climate that are less likely to be found in the classroom environment. The item difficulties for all the items in this scale are all negative, which means that the items are relatively easier to endorse.

vi. Teacher-student relationship

Five items on teacher-student relations are included from the PISA 2012 Student Questionnaire. This scale provides information about the perceptions students have of their teachers’ interest in their performance. There are four response categories varying from “Strongly agree”, “Agree”, “Disagree” to “Strongly disagree”. All items must be reversed. The statement that students generally find the most difficult to agree on is that most of their teachers really listen to what students have to say.

vii. Construction of variables for the Reading Profile

Questions ST025Q01 to ST025Q05 included from PISA 2009 Student Questionnaire about students’ frequency of reading certain materials must be recoded into five dummy variables named “magazine”, “comic”, “fiction”, “non-fiction” and “news”. Each of these dummy variables equals one if the student declares reading this material “several times a month” or “several times a week” and zero if the student declares reading this material “never or almost never”, “a few times a year” or “about once a month”.

Two other variables must be derived from questions relative to two meta-cognition tasks, “Understanding and remembering” (UNDREM) and “Summarising” (METASUM). Both meta-cognition tasks consist of a stem (which is a reading task) and a set of strategies.

For each strategy listed in items ST041Q01 to ST041Q06 from the PISA 2009 Student Questionnaire, students are asked to rate the usefulness of the strategy for understanding and remembering a text, using a scale ranging from 1 to 6. For each student, nine scores are calculated according to the following set of nine order relations:

- If ST041Q03>ST041Q01 then student’s score is set to 1, else 0.
- If ST041Q03>ST041Q02 then student’s score is set to 1, else 0.
- If ST041Q03>ST041Q06 then student’s score is set to 1, else 0.
- If ST041Q04>ST041Q01 then student’s score is set to 1, else 0.

- If ST041Q04>ST041Q02 then student's score is set to 1, else 0.
- If ST041Q04>ST041Q06 then student's score is set to 1, else 0.
- If ST041Q05>ST041Q01 then student's score is set to 1, else 0.
- If ST041Q05>ST041Q02 then student's score is set to 1, else 0.
- If ST041Q05>ST041Q06 then student's score is set to 1, else 0.

If any of the rates used in an order relation is missing, then the associated score must also be missing. The nine created scores must be added together and divided by nine. The resulting proportion is the student's score on the UNDREM scale.

For each strategy listed in items ST042Q01 to ST042Q05 from the PISA 2009 Student Questionnaire, students are asked to rate the usefulness of the strategy for *summarising* a text, using a scale ranging from 1 to 6. For each student, eight scores must be calculated according to the following set of eight order relations:

- If ST042Q04>ST042Q01 then student's score is set to 1, else 0.
- If ST042Q04>ST042Q03 then student's score is set to 1, else 0.
- If ST042Q04>ST042Q02 then student's score is set to 1, else 0.
- If ST042Q05>ST042Q01 then student's score is set to 1, else 0.
- If ST042Q05>ST042Q03 then student's score is set to 1, else 0.
- If ST042Q05>ST042Q02 then student's score is set to 1, else 0.
- If ST042Q01>ST042Q02 then student's score is set to 1, else 0.
- If ST042Q03>ST042Q02 then student's score is set to 1, else 0.

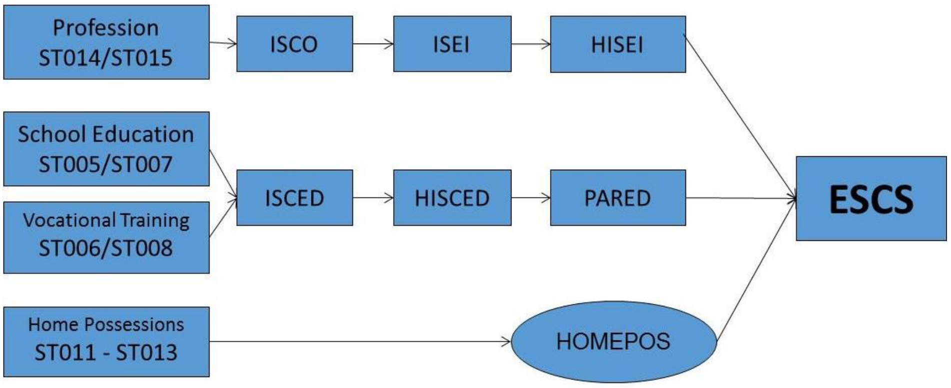
If any of the rates used in an order relation is missing, then the associated score must also be missing. The eight created scores must be added together and divided by eight. The resulting proportion is the student's score on the METASUM scale.

e. Other complex questionnaire variables

i. Index of economic, social and cultural status (ESCS)

One key variable for the analyses reported in the School Report is the index of economic, social and cultural status (ESCS). The ESCS index was used first in the PISA 2000 analysis, and at that time, was derived from five indices: highest occupational status of parents (HISEI), highest educational level of parents (PARED), and three IRT scales based on student reports on home possessions: family wealth (WEALTH), cultural possessions (CULTPOSS) and home educational resources (HEDRES). Since PISA 2003, the ESCS has been derived from three indices: highest parental occupation (HISEI), highest educational level of parents (PARED), and one IRT scale based on student reports on home possessions, including books in the home (HOMEPOS). However, until PISA 2012, the PCA was based on OECD countries only. In PISA 2015, the PCA is estimated across all countries concurrently. Thus, all countries and economies contribute equally to the estimation of ESCS scores.

Figure 2. Computation of ESCS in PISA 2015



Missing values for students with missing data for only one variable must be imputed with predicted values plus a random component based on a regression on the other two variables. If there are missing data on more than one variable, ESCS is not computed for that case and a missing value is assigned for ESCS. The imputed variables were standardised for OECD countries and partner countries/economies with an OECD mean of 0 and a standard deviation of 1.

In the PBTS, ESCS scores are obtained as:

$$\mathrm{ESCS} = \frac{\mathrm{HISEI}' + \mathrm{PARED}' + \mathrm{HOMEPOS}'}{3}$$

where HISEI' PARED' and HOMEPOS' are the "OECD-standardised" variables, and. HISEI must be standardised with a mean of 51.50 and a standard deviation of 21.98, PARED with a mean of 13.85 and a standard deviation of 3.08, and HOMEPOS with a mean of 0.00 and a standard deviation of 1.00.

ii. Highest occupational status of parents (HISEI)

Occupational data for both the student's father and mother are obtained by asking open-ended questions in the Student Questionnaire. The responses must be coded to four-digit International Standard Classification of Occupation (ISCO) codes (International Labour Organization (ILO), 2007[21]). Details about ISCO codes can be found in the PISA 2012 Student Questionnaire Codebook (ACER, 2015[22]) that will be provided by the OECD. Once the parents' occupations are coded, they must then be mapped to the international socio-economic index of occupational status (ISEI) (Ganzeboom, 2010[23]). In PISA 2015, the new ISCO and ISEI in their 2008 version were used, rather than the 1988 versions that had been applied in the previous four cycles.

Three indices are calculated based on this information: father's occupational status (BFMJ2); mother's occupational status (BMMJ1); and the highest occupational status of parents (HISEI), which corresponds to the higher ISEI score of either parent or to the only available parent's ISEI score. For all three indices, higher scores indicate higher levels of occupational status.

iii. Education levels of parents (PARED)

Students' responses regarding parental education must be classified using ISCED (UNESCO Institute for Statistics, 2012[24]). Indices on parental education must be constructed by recoding educational qualifications into the following categories: (0) None, (1) ISCED 1 (primary education), (2) ISCED 2 (lower secondary),

(3) ISCED Level 3B or 3C (vocational/pre-vocational upper secondary), (4) ISCED 3A (general upper secondary) and/or ISCED 4 (non-tertiary post-secondary), (5) ISCED 5B (vocational tertiary) and (6) ISCED 5A or 6 (theoretically oriented tertiary and post-graduate). Indices with these categories must be created for the students' mother (MISCED) and the students' father (FISCED). In addition, the index on the highest educational level of parents (HISCED) corresponds to the higher ISCED level of either parent. The index for highest educational level of parents must also be recoded into estimated number of years of schooling (PARED). The mapping of ISCED levels to years of schooling (PARED) was updated in 2009 and 2015 for some countries, taking into account changes in countries' educational systems. The most recent mapping of ISCED levels to years of schooling used in PISA is adopted in the analysis of PBTS data[4] (OECD, 2018$_{[25]}$).

iv. Household possessions

In PISA 2015, students reported the availability of 16 household items at home (ST011), including three country-specific household items that were seen as appropriate measures of family wealth within the context of the country. In addition, students reported the amount of possessions and books at home (ST012, ST013). HOMEPOS is a summary index of all household and possessions (ST011, ST012 and ST013). HOMEPOS is also one of three components in the construction of the PISA index of economic, social and cultural status. In PBTS, the 16 household items are included in the assessment.

The computation of the home possessions scale for PISA 2015 was performed in a way that differed from previous cycles. The IRT model has changed from the Partial Credit model to the Generalised Partial Credit Model for the purpose of cross-cultural comparability (see Box 3 for further explanation). Categories for the number of books in the home are unchanged in PISA 2015. The variable indicating the number of books at home is recoded from the original 6 categories into 3: (0) 0-25 books, (1) 26-100 books, (2) more than 100. Questions 15-ST012Q01 and 15-ST012Q02 are recoded from the original 4 categories into 3: (0) "None or one", (1) "Two" (2) "Three or more". The rest of the questions retain four categories. The ST011 items (1="yes", 2="no") were reverse coded, so that a higher level indicates the presence of the indicator. Table 4 shows the wording of items used for the computation of HOMEPOS.

HOMEPOS is also one of three components in the construction of the index of economic, social and cultural status (or ESCS; see the subsection on ESCS index construction above in this section).

[4] Currently, the PBTS uses the PISA 2018 mapping of ISCED.

Table 4. Home possession items

ST011	In your home, do you have:
ST011Q01TA	A desk to study at
ST011Q02TA	A room of your own
ST011Q03TA	A quiet place to study
ST011Q04TA	A computer you can use for school work
ST011Q05TA	Educational software
ST011Q06TA	A link to the Internet
ST011Q07TA	Classic literature (e.g. <Shakespeare>)
ST011Q08TA	Books of poetry
ST011Q09TA	Works of art (e.g. paintings)
ST011Q10TA	Books to help with your school work
ST011Q11TA	<Technical reference books>
ST011Q12TA	A dictionary
ST011Q16NA	Books on art, music, or design
ST011Q17TA	<Country-specific wealth item 1>
ST011Q18TA	<Country-specific wealth item 2>
ST011Q19TA	<Country-specific wealth item 3>
ST12	**How many of these are there in your home?**
ST012Q01TA	Televisions
ST012Q02TA	Cars
ST012Q03TA	Rooms with a bath or shower
ST012Q05NA	<Cell phones> with Internet access (e.g. smartphones)
ST012Q06NA	Computers (desktop computer, portable laptop, or notebook)
ST012Q07NA	<Tablet computers> (e.g. <iPad®>, <BlackBerry® PlayBook™>)
ST012Q08NA	E-book readers (e.g. <Kindle™>, <Kobo>, <Bookeen>)
ST012Q09NA	Musical instruments (e.g., guitar, piano)
ST013Q01TA	**How many books are there in your home?**

Source: OECD (2017), *PISA 2015 Technical Report*, Chapter 16

Box 3. Scaling procedures of Context Questionnaire in PISA 2015

As in previous cycles of PISA, one subset of the derived variables was constructed using IRT (item response theory) scaling methodology. In the IRT framework, a number of different models can be distinguished, the Generalised Partial Credit Model (GPCM) (see below) was used for constructing derived variables in the PISA 2015 Context Questionnaires.

For each item, item responses are modelled as a function of the latent construct, *θ_j*. With the one-parameter model (Rasch, 1960[26]) for dichotomous items, the probability of person *j* selecting category 1 instead of 0 is modelled as

$$P\left(X_{ji} = 1|\theta_j, \beta_i\right) = \frac{exp\left(\theta_j - \beta_i\right)}{1 + exp\left(\theta_j - \beta_i\right)}$$

where *P*(*X_ji*=1) is the probability of person *j* to score 1 on item *i*, *θ_j* is the estimated latent trait of person *j* and β_i the estimated location or difficulty of item *i* on this dimension. In the case of items with more than two (*m*) categories (e.g. Likert-type items), this model can be generalised to the Partial Credit Model (Masters, 1982[27]), which takes the form of

$$P\left(X_{ji} = k|\theta_j, \beta_i, \boldsymbol{d_i}\right) = \frac{exp\left(\sum_{r=0}^{k} \theta_j - \beta_i + d_{ir}\right)}{\sum_{u=0}^{m_i} exp\left(\sum_{r=0}^{u} \theta_j - \beta_i + d_{ir}\right)}$$

where $P(X_{ji}=k)$ denotes the probability of person *j* to score *k* on item *i* out of the m_i possible scores on the item. *θ_j* denotes the person's latent trait, the item parameter β_i gives the general location or difficulty of the item on the latent continuum and d_{ir} denote additional step parameters. This model has been used throughout previous cycles of PISA for scaling derived variables of the context questionnaires. However, research literature (especially: (Glass and Jehangir, 2014[28]) suggests that a generalisation of this model, the GCPM (Muraki, 1992[29]), is more appropriate in the context of PISA, since it allows the item discrimination to vary between items within any given scale. This model takes the form of

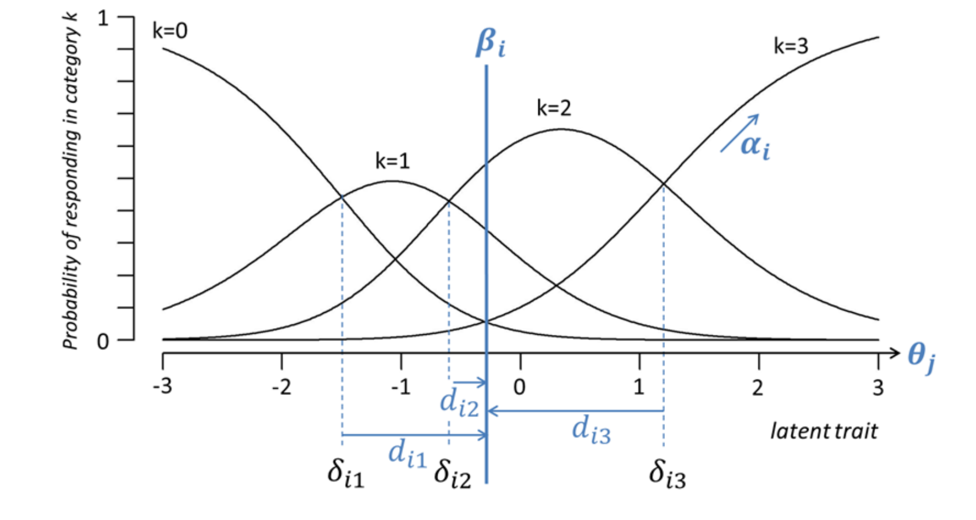
$$P\left(X_{pi} = k|\theta_j, \beta_i, \alpha_i, d_i\right) = \frac{exp\left(\sum_{r=0}^{k} \alpha_i\left(\theta_j - \beta_i + d_{ir}\right)\right)}{\sum_{u=0}^{m_i} exp\left(\sum_{r=0}^{u} \alpha_i\left(\theta_j - \beta_i + d_{ir}\right)\right)}$$

in which the additional discrimination parameter α_i allows for the items of a scale to contribute with different weights to the measurement of the latent construct.

Most of the scales were analysed based on 2015 data only (regular scales) and others, mostly science-related scales were analysed to allow for comparisons with the weighted likelihood estimates (WLEs) (Warm, 1989[30]) obtained in PISA 2006 (trend scales, see below).

Box 3. Scaling procedures of Context Questionnaire in PISA 2015 (cont)

The GCPM described above contains three kinds of item parameters: one relating to the general location or difficulty of the item (β), one relating to the deviance of each of the single response categories from this location parameter (d), and one relating to the item's discrimination or slope (a). The following figure displays the category characteristic curves (CCC) of a four-category item (e.g. a Likert-type item with response categories "Strongly disagree", "Disagree", "Agree", and "Strongly agree"). The three kinds of GPCM item parameters were included in this representation. The overall item location or difficulty parameter, *β*, can be regarded as the item's location on the latent continuum of the construct to be measured. The *m*-1 threshold parameters, *d*, of an *m*-category item represent deviations from this general location.



Item characteristic curves for a four-category item under the GPCM. Model parameters are highlighted in blue.

Source: OECD (2017), *PISA 2015 Technical Report*, Chapter 16.

f. Social and Emotional Skills

PBTS uses 40 items from the Survey on Social and Emotional Skills (SSES), which explore five factors. The five factors are assertiveness, curiosity, empathy, optimism, and self-control. The item assignment is shown in Table 5. Each item is scored from 0 to 4 for items with positively worded statements and reverse-scored for the negatively worded items. In PBTS, scores of the five factors are calculated for each scale separately. Student responses to items are treated as polytomous data and the GPCM is applied for the scaling. $N(0, 1)$ is assumed for scores of a group; therefore, it should be noted that the scores between the two different testing cycles are not able to be compared in PBTS.

Table 5. 40 items used in PBTS SES module

Scale	Item	ItemID	Reverse (0=no, 1= yes)
Assertiveness	A leader	STA1401	0
Assertiveness	Want to be in charge	STA1402	0
Assertiveness	Know how to convince others to do what I want	STA1403	0
Assertiveness	Enjoy leading others	STA1404	0
Assertiveness	Dislike leading a team	STA1406	1
Assertiveness	Like to be a leader in my class	STA1407	0
Assertiveness	Like to be the leader of a group	STA1408	0
Assertiveness	Dominant, and act as a leader	STA1409	0
Curiosity	Curious about many different things	STA0401	0
Curiosity	Eager to learn	STA0402	0
Curiosity	Like to ask questions	STA0403	0
Curiosity	Like to know how things work	STA0404	0
Curiosity	Like learning new things	STA0405	0
Curiosity	Don't like learning	STA0406	1
Curiosity	Love learning new things in school	STA0407	0
Curiosity	Find science interesting	STA0408	0
Empathy	Helpful and unselfish with others	STA0502	0
Empathy	Important to me that my friends are okay	STA0503	0
Empathy	Can sense how others feel	STA0504	0
Empathy	Know how to comfort others	STA0505	0
Empathy	Predict the needs of others	STA0506	0
Empathy	Understand what others want	STA0507	0
Empathy	Warm toward others	STA0508	0
Empathy	Rarely ask others how they are feeling	STA0509	1
Optimism	Often feel sad	STA0901	1
Optimism	Believe good things will happen to me	STA0902	0
Optimism	Wake up happy almost every day	STA0903	0
Optimism	Always positive about the future	STA0904	0
Optimism	Enjoy life	STA0905	0
Optimism	Look at the bright side of life	STA0907	0
Optimism	A happy person	STA0908	0
Optimism	Expect bad things to happen	STA0910	1
Self-control	Careful with what I say to others	STA1701	0
Self-control	Can control my actions	STA1702	0
Self-control	Think carefully before doing something	STA1703	0
Self-control	Avoid mistakes by working carefully	STA1704	0
Self-control	Say the first thing that comes to my mind	STA1706	0
Self-control	Like to make sure there are no mistakes	STA1707	0
Self-control	Stop to think before acting	STA1708	0
Self-control	Often rush into action without thinking	STA1709	1

3.4 Survey weighting (the balanced repeated replication method)

Survey weights are required to facilitate calculation of appropriate estimates of sampling error and making valid estimates and inferences of the population. OECD statisticians must calculate survey weights for all assessed students, to get estimates of standard errors, conduct significance tests and create confidence intervals appropriately.

A replication methodology must be employed to estimate the sampling variances of PBTS parameter estimates. This methodology accounts for the variance in estimates due to the sampling of students. Additional variance due to the use of plausible values from the posterior distributions of scaled scores is captured separately as measurement error.

a. The balanced repeated replication method

The approach used for calculating sampling variances for PISA, and therefore PBTS, estimates is known as balanced repeated replication (BRR), or balanced half-samples. The particular variant known as Fay's method must be used[5]. This variant of the BRR method must be implemented as follows.

In each school, students must be ranked on the basis of their first principal component that was used to generate the plausible values. The first two students in the ordered list are paired; the following two students are paired and so on. If a school has an odd number of students, the last three students are grouped to form a triple. Let us assume that there are H pairs of students, also referred as variance strata or zones, or pseudo-strata in the sampling literature. Pairs are numbered sequentially, 1 to H.

A set of 80 replicate weights must then be created. Each of these replicate weights is formed by assigning a weight of 1.5 to one of the two students in the stratum, and a weight of 0.5 to the remaining student. In cases where there are three units in a triple, either one of the students (designated at random) receive a factor of 1.7071 for a given replicate, with the other two students receiving factors of 0.6464, or else the one student receives a factor of 0.2929, and the other two students receive factors of 1.3536. The determination as to which students receive inflated weights and which receive deflated weights is carried out in a systematic fashion, based on the entries of the first H rows in a Hadamard matrix[6] of order 80. A Hadamard matrix contains entries that are +1 and –1 in value.

For pairs of students:

- The +1 in the Hadamard matrix is converted to a weight of 1.5 for the first student of the pair, and 0.5 for the second student of the pair;
- The -1 in the Hadamard matrix is converted to a weight of 0.5 for the first student of the pair, and 1.5 for the second student of the pair.

For triples of students:

- The +1 in the Hadamard matrix is converted to a weight of 1.7071 for the first student of the pair, and 0.6464 for the other two students of the triple;
- The -1 in the Hadamard matrix is converted to a weight of 0.2929 for the first student of the pair, and 1.3536 for the other two students of the triple.

Table 6 and Table 7 describe how the replicate weights are generated for this method in a fictitious example where there are 21 students in a participating school. Table 6 displays an example of a Hadamard matrix. Table 7 shows how the replicate weights are assigned to each student.

[5] This method is similar in nature to the Jackknife method used in other international studies of educational achievement.

[6] Details concerning Hadamard matrices are given in (Wolter, 2007[33]).

Table 6. Hadamard Matrix

	Column 1	Column 2	Column 3	…	Column 80
Row 1	1	1	1		1
Row 2	1	1	-1		-1
Row 3	1	-1	1		1
Row 4	1	1	-1		1
Row 5	1	1	1		1
Row 6	1	1	1		-1
Row 7	1	-1	1		-1
Row 8	1	-1	-1		-1
Row 9	-1	-1	-1		1
Row 10	1	1	-1		-1

Table 7. Replicates for the Balanced Replicate method

Pseudo-stratum	Student identifier	FULL WEIGHT	BRR WEIGHT 1	BRR WEIGHT 2	BRR WEIGHT 3	…	BRR WEIGHT 80
1	1	1	1.5	1.5	1.5		1.5
1	2	1	0.5	0.5	0.5		0.5
2	3	1	1.5	1.5	0.5		0.5
2	4	1	0.5	0.5	1.5		1.5
3	5	1	1.5	0.5	1.5		1.5
3	6	1	0.5	1.5	0.5		0.5
4	7	1	1.5	1.5	0.5		1.5
4	8	1	0.5	0.5	1.5		0.5
5	9	1	1.5	1.5	1.5		1.5
5	10	1	0.5	0.5	0.5		0.5
6	11	1	1.5	1.5	1.5		0.5
6	12	1	0.5	0.5	0.5		1.5
7	13	1	1.5	0.5	1.5		0.5
7	14	1	0.5	1.5	0.5		1.5
8	15	1	1.5	0.5	0.5		0.5
8	16	1	0.5	1.5	1.5		1.5
9	17	1	0.5	0.5	0.5		1.5
9	18	1	1.5	1.5	1.5		0.5
10	19	1	0.7071	0.7071	0.2929		0.2929
10	20	1	0.6464	0.6464	1.3536		1.3536
10	21	1	0.6464	0.6464	1.3536		1.3536

Source: Based on OECD (2009*), PISA Data Analysis Manual. Second edition*, Table 4.12/4.13.

b. The sampling variance estimator

As with all replication methods, the statistic of interest (φ) is computed based on the sample ($\widehat{\varphi}$), and then again on each replicate ($\widehat{\varphi}_{(i)}, i = 1 \text{ to } 80$). The replicates are then compared to the whole sample estimate to get the sampling variance, as follows:

$$\sigma^2_{(\widehat{\varphi})} = \frac{1}{20}\sum_{i=1}^{80}(\widehat{\varphi}_{(i)} - \widehat{\varphi})^2$$

3.5 Statistical procedures for generating school report results

This section of the report has been developed to provide statisticians with the techniques needed to correctly analyse the PBTS database. It helps them to confidently and accurately implement procedures used for the production of the PBTS School Reports. This section is largely based on the procedures described in the *PISA Data Analysis Manual* (OECD, 2009[31]) ([32]) which the reader can refer to for further detail. This section will cover all cases of computations required for generating School Report results.

a. Univariate statistics for context variables

Replicate weights have to be used for the computation of the standard error for any population estimate. The standard error of statistics for context variables only (i.e., not involving performance variables) simply equals the sampling error, i.e., the square root of the sampling variance estimator reported in Section b.

i. Percentage of students per category

For categorical variables, the statistic of interest is usually a percentage of students per category. Each percentage is estimated after deleting cases with missing values for the variable of interest in the dataset. The procedure for estimating the percentage of students per category and the corresponding standard error requires first computing the percentage for the original sample (without weighting student data), and then computing 80 other percentages, each of them by weighting the student sample with one of the 80 replicates.

SAS® and SPSS® syntaxes and macros (respectively PROC_FREQ_NO_PV.SAS and MCR_SE_GRPPCT.SPS) for computing the percentages and their standard errors are presented in chapter 7 of the *PISA Data Analysis Manual* (OECD, 2009[31]) ([32]).

For reporting, percentages of students:

- reporting that disciplinary issues in <test language> lessons occur “never or hardly ever” or “in some lessons”;
- reporting that disciplinary issues in mathematics lessons occur “never or hardly ever” or “in some lessons”;
- who agree or strongly agree with statements regarding teacher-student relations at school;
- per reading profile;
- who feel either confident or very confident about having to do various mathematics tasks;
- who feel either confident or very confident about having to do various science tasks;

- who agree or strongly agree with statements regarding their instrumental motivation in mathematics;
- who agree or strongly agree with statements regarding their instrumental motivation in science;

as well as their standard errors must be computed.

ii. School mean indices

To compute a school mean index and its corresponding standard error, it is also necessary to first compute the mean on the original student sample, and then to compute 80 other means, each of them by weighting the data with one of the 80 replicates.

SAS® and SPSS® syntaxes and macros (respectively PROCMEAN_NO_PV.SAS and MCR_SE_UNIV.SPS) for computing the percentages and their standard errors are presented in chapter 7 of the *PISA Data Analysis Manual* (OECD, 2009[31]) ([32]).

For reporting, the school mean of the economic, social and cultural status index (ESCS) as well as its standard errors must be computed.

iii. Identification of "similar" schools in regards to socio-economic background

Several figures in the School Report compare the school's results with those of similar schools with regards to socio-economic background. "Similar" schools are schools that participated in PISA 2012 and whose ESCS is within +/- 0.25 of the ESCS index of the school having participated in the PBTS.

b. Students' average scores and their standard errors

As described above, the cognitive data in the PBTS are scaled with the mixed coefficients multinomial logit model and the performance of students is denoted with plausible values (PVs). For each domain, five plausible values per student are included in the international databases. This section describes how to perform analyses with plausible values, so it is useful when reporting results on student performances and their relationships with student or school characteristics.

The computation of a statistic with plausible values always consists of six steps, regardless of the required statistic.

1. The required statistic has to be computed for each plausible value. Given the BRR design, it follows that 81 estimates are necessary to get the final estimate and its standard error. Therefore, any analysis that involves five plausible values will require 405 estimates (5 × 81). To estimate a mean score and its respective standard error, 405 means must be calculated. The means estimated on the original sample (without weighting) are denoted $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_4$ and $\hat{\mu}_5$. From the 80 replicates applied on each of the five plausible values, five sampling variances are estimated, denoted respectively $\sigma^2_{(\hat{\mu}_1)}$,$\sigma^2_{(\hat{\mu}_2)}$,$\sigma^2_{(\hat{\mu}_3)}$, $\sigma^2_{(\hat{\mu}_4)}$ and $\sigma^2_{(\hat{\mu}_5)}$. These five mean estimates and their respective sampling variances are provided in Table 8.

Table 8. The 405 mean estimates

Weight	PV1	PV2	PV3	PV4	PV5
None	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
Replicate 1	$\hat{\mu}_{1\ 1}$	$\hat{\mu}_{2\ 1}$	$\hat{\mu}_{3\ 1}$	$\hat{\mu}_{4\ 1}$	$\hat{\mu}_{5\ 1}$
Replicate 2	$\hat{\mu}_{1\ 2}$	$\hat{\mu}_{2\ 2}$	$\hat{\mu}_{3\ 2}$	$\hat{\mu}_{4\ 2}$	$\hat{\mu}_{5\ 2}$
Replicate 3	$\hat{\mu}_{1\_3}$	$\hat{\mu}_{2\_3}$	$\hat{\mu}_{3\_3}$	$\hat{\mu}_{4\_3}$	$\hat{\mu}_{5\_3}$
….	…	….	…	…	…
Replicate 80	$\hat{\mu}_{1\ 80}$	$\hat{\mu}_{2\ 80}$	$\hat{\mu}_{3\ 80}$	$\hat{\mu}_{4\ 80}$	$\hat{\mu}_{5\ 80}$
Sampling variance	$\sigma^2_{(\hat{\mu}_1)}$	$\sigma^2_{(\hat{\mu}_2)}$	$\sigma^2_{(\hat{\mu}_3)}$	$\sigma^2_{(\hat{\mu}_4)}$	$\sigma^2_{(\hat{\mu}_5)}$

Source: OECD (2009), PISA Data Analysis Manual, Second edition, Table 8.1.

2. The final mean estimate is equal to the average of the five mean estimates, i.e.,

$$\hat{\mu} = \frac{1}{5}(\hat{\mu}_1 + \hat{\mu}_2 + \hat{\mu}_3 + \hat{\mu}_4 + \hat{\mu}_5).$$

3. The final sampling variance is equal to the average of the five sampling variances, i.e.,

$$\hat{\sigma}^2_{\text{sampling}} = \frac{1}{5}(\sigma^2_{(\hat{\mu}_1)} + \sigma^2_{(\hat{\mu}_2)} + \sigma^2_{(\hat{\mu}_3)} + \sigma^2_{(\hat{\mu}_4)} + \sigma^2_{(\hat{\mu}_5)}).$$

4. The imputation variance, also denoted measurement error variance, is computed as

$$\hat{\sigma}^2_{\text{measure}} = \frac{1}{4}\sum_{i=1}^{5}(\hat{\mu} - \hat{\mu}_i)^2$$

5. Two types of standard errors can be computed depending on the reporting purpose. Whenever a mean score is compared with a specific PISA result or put in parallel of a specific PISA scale, the appropriate link error must be included in the calculation of the standard error of the mean estimate. For instance, if the school average score of the PBTS mathematics scale is compared with the country's average score at PISA 2015 mathematics test, the link error between the PBTS and the PISA 2015 mathematics tests must be included in the estimation of the standard error on the school's average score. The total error variance results from the combination of the sampling error variance, the measurement error variance and the linking error variance (denoted $\hat{\sigma}^2_{\text{link}}$), and it is computed as:

$$\hat{\sigma}^2_{\text{final with link error}} = 1.2\,\hat{\sigma}^2_{\text{measure}} + \hat{\sigma}^2_{\text{sampling}} + \hat{\sigma}^2_{\text{link}}.$$

   Whenever a mean estimate (such as the school average score) is presented as such, i.e., alone, then no link error must be taken into account in the calculation of the standard error. The sampling variance and the imputation variance are combined to obtain a final error variance as:

$$\hat{\sigma}^2_{\text{final without linking}} = 1.2\,\hat{\sigma}^2_{\text{measure}} + \hat{\sigma}^2_{\text{sampling}}.$$

6. The standard error equals the square root of the total error variance.

References

ACER (2015), *Codebook for PISA 2012 Main Study Student Questionnaire*, https://www.oecd.org/pisa/pisaproducts/PISA12_stu_codebook.pdf. [22]

Asparouhov, T. and B. Muthén (2014), "Multiple-Group Factor Analysis Alignment", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 21/4, pp. 495-508, https://doi.org/10.1080/10705511.2014.919210. [16]

Ganzeboom, H. (2010), *Questions and answers about ISEI–08*, http://www.harryganzeboom.nl/isco08/qa-isei-08.htm. [23]

Glass, C. and K. Jehangir (2014), *Modeling country-specific differential item functioning, Ana*, In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Springer. [28]

International Labour Organization (ILO) (2007), *Resolution concerning updating the International Standard Classification of Occupations*, http://www.ilo.org/public/english/bureau/stat/isco/docs/resol08.pdf (accessed on 17 March 2015). [21]

International Test Commission (2013), *ITC Guidelines on Test Use*, https://www.intestcom.org/files/guideline_test_use.pdf (accessed on 6 January 2017). [18]

Masters, G. (1982), *A Rasch model for partial credit scoring*, Psychometrik. [27]

Mislevy, R. (1985), *Estimation of latent group effects*, Journal of the American Statistical Association.

Muraki, E. (1992), *A generalized partial credit model: Application of an EM algorithm*. *Psychological Measurement*, No. 16, pp. 159-176. [29]

OECD (2020), *PISA 2018 Technical Report*, https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TecReport-Ch-04-Sample-Design.pdf. [13]

OECD (2018), *PISA 2018 Mapping of ISCED levels to years*, https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TechRep-Final-AnnexD.xlsx. [25]

OECD (2017), *PISA 2015 Technical Report*, http://www.oecd.org/pisa/data/2015-technical-report/. [17]

OECD (2014), *PISA 2012 Technical Report*, http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf. [19]

OECD (2012), *PISA 2009 Technical Report*, OECD, Paris, http://www.oecd.org/pisa/pisaproducts/50036771.pdf.

OECD (2009), *PISA 2006 Technical Report*, OECD, Paris, http://www.oecd.org/pisa/pisaproducts/42025182.pdf.

OECD (2009), *PISA Data Analysis Manual: SAS®, Second Edition*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264056251-en. [31]

OECD (2009), *PISA Data Analysis Manual: SPSS®, Second Edition*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264056275-en. [32]

Rasch, G. (1960), *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*, Nielsen & Lydiche. [26]

UNESCO Institute for Statistics (2012), *International Standard Classification of Education, ISCED 2011*, http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf. [24]

von Davier, M. (2008), *A general diagnostic model applied to language testing data*, British Journal of Mathematical and Statistical Psychology. [20]

Warm, T. (1989), *Weighted Likelihood Estimation of Ability in Item Response Theory*, Psychometrika, No. 54. [30]

Wolter, K. (2007), *Introduction to Variance Estimation*, Second edition, Springer. [33]

Wu, M.L., Adams, R.J., & Wilson, M.R. (2007), *ACER ConQuest version 2.0: generalised item response modelling software*, Camberwell, VIC, ACER Press.