OECD Education Working Papers No. 126

PISA For Development Technical Strand A: Enhancement of PISA Cognitive Instruments

Ray Adams,

John Cresswell

https://dx.doi.org/10.1787/5jm5fb3f85s0-en

Unclassified **EDU/WKP(2015)15**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

15-Dec-2015

English - Or. English

DIRECTORATE FOR EDUCATION AND SKILLS

EDU/WKP(2015)15
Unclassified

PISA FOR DEVELOPMENT TECHNICAL STRAND A: ENHANCEMENT OF PISA COGNITIVE INSTRUMENTS

Education Working Paper 126

by Ray Adams and John Cresswell, The Australian Council for Educational Research

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Ray Adams, The Australian Council for Educational Research (adams@acer.edu.au)
Michael Ward, Senior Policy Analyst, DCD/GPP (Michael.Ward@oecd.org)

JT03388327

Complete document available on OLIS in its original format

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.

English - Or. English

OECD EDUCATION WORKING PAPERS SERIES

OECD Working Papers should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed herein are those of the author(s).

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcome, and may be sent to the Directorate for Education and Skills, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France.

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.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

The names of countries and territories used in this joint publication follow the practice of the OECD.

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

----

www.oecd.org/edu/workingpapers

PISA FOR DEVELOPMENT
TECHNICAL STRAND A: ENHANCEMENT OF PISA COGNITIVE INSTRUMENTS

Abstract

The OECD is planning to enhance existing PISA assessment instruments in reading, mathematics and science so that they will be more suitable to the context of developing countries. The main purpose of this paper is to identify the main technical issues associated with this aim. The paper reports detailed analysis of the existing PISA item pool and its suitability for countries which have students of average limited capacity. The paper cautions that the fit of developing country data to the PISA model is not good and that modifications to address some of the deviations should be explored. The use of learning metrics to describe dimensions of educational progression is at the core of the PISA reporting methodology and requires a consistency across countries in item behaviour that is not apparent for developing countries. The paper recommends that any process to move towards enhancing the instruments must be undertaken with extensive consultation with the countries involved.

Résumé

L'OCDE prévoit le renforcement des instruments d'évaluation PISA existants en compréhension de l'écrit, en mathématiques et en sciences, afin de mieux les adapter au contexte des pays en développement. Ce document vise avant tout à mettre au jour les principaux enjeux techniques associés à ce projet. Il présente une analyse détaillée de l'ensemble des items PISA existants et de leur pertinence pour des pays où les élèves ont un niveau de compétences moyen limité. Il met en évidence la mauvaise adaptation des données des pays en développement au modèle PISA et suggère l'exploration de modifications afin de pallier certains des écarts. L'utilisation de mesures de l'apprentissage pour décrire les dimensions de la progression en matière d'éducation est au cœur de la méthodologie de PISA et nécessite une certaine cohérence entre les pays en termes de comportement des items qui n'est pas apparente pour les pays en développement. Ce document recommande que tout processus en vue du renforcement des instruments d'évaluation s'accompagne d'une consultation approfondie auprès des pays concernés.

TABLE OF CONTENTS

Tables

Figures

ACRONYMS

ACER	Australian Council for Educational Research
EGRA	Early Grade Reading Assessment
ICCS	International Civics and Citizenship Study
IEA	International Association for the Evaluation of Educational Achievement
IRT	Item Response Theory - relating to scoring student answers in a test
LLECE	Laboratory for Assessment of the Quality of Education
OECD	Organisation for Economic Cooperation and Development
PASEC	*Programme d'Analyse des Systèmes Educatifs de la CONFEMEN* (Programme for the Analysis of Educational Systems of the CONFEMEN countries)
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
RASCH	George Rasch devised a model for scoring student responses to an assessment
SABER	Systems Approach for Better Education Results (World Bank project)
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
TIMSS	Trends in Mathematics and Science Study
UNESCO	United Nations Educational Social and Cultural Organisation
UNICEF	United Nations Children's Fund

PREAMBLE

The success of PISA since its first survey administration in 2000 has attracted an ever increasing number of participating countries. The countries benefit by comparing the capacity of their student body with other countries throughout the world.

PISA currently builds on the experience of over 70 countries, among them several low- and middle-income countries. The original framework for PISA was based on OECD countries, but as the membership in PISA expanded, the frameworks evolved as well. All countries that take part in a PISA assessment also participate in the framework development. However, with an increasingly diverse group of countries, the targeting of the assessment may be less appropriate for lower-middle-income countries and low-income countries. The PISA for Development project in coordination with the regular PISA surveys will allow better targeting of the assessments.

Countries are participating in the PISA for Development initiative with the expectation that they will have access to a more precise measure of their students' capacity and through an appropriately targeted questionnaire will better understand the factors which contribute to student success in their own country.

At the same time they expect that the results of the PISA for Development assessments to be fully comparable to the results of all countries which undertake PISA. As such the principles underlying the participation must be the same as for PISA and countries undertaking the assessment will do so in full accordance with the agreed PISA Technical Standards which apply to test design, administration, data entry, translation and sampling.

For the current project it has been decided that no new cognitive items will be developed and as a consequence PISA for Development will not be able to enhance the current descriptions of proficiency levels. There is, however, a large set of existing secure PISA items which can be included in an appropriate test design targeted to give maximum information to the set of countries participating. PISA for Development participants have expressed a desire to review the secure pool with a view to selecting items based upon their cultural and contextual suitability.

The targeted test which will be administered to students of countries participating in the PISA for Development project should give a precise measure of the country's strengths and weaknesses so that policy makers and practitioners can implement changes which will lead to long term improvement in student capacity. This approach of testing the students' capacity now and then planning for future improvements is preferable to having a very difficult test which assesses where it is thought the students should be.

While it is recognised that it is important to better target the PISA for Development instruments participants have also expressed the importance of a test that assesses across the full range of the PISA scale.

EXECUTIVE SUMMARY

The main purpose of this work is to identify the main technical issues to be addressed by the PISA for Development project which aims to enhance the descriptive power of PISA's cognitive assessments in reading, mathematics and science, especially in developing countries. This work on the cognitive instruments of PISA complements separate papers that are being prepared concurrently for technical Strand B of PISA for Development (contextual questionnaires) and Technical Strand C (approaches to out-of-school children).

Main observations

Observation 1: In any move to expand the use of PISA to a greater number of countries it would be essential to carry out a complete review of the assessment frameworks in consultation with those countries. It might be expected that the areas currently included for assessment which are seen as priorities by high-income countries may not coincide with the areas that are seen as priorities for low-income and middle-income countries. At the same time any extension of the framework will need to continue to incorporate the original.

Observation 2: The PISA standard of testing in the language of instruction should be reviewed for PISA for Development since there may be large numbers of students for which an assessment in the official language of instruction does not provide a valid indication of a student's proficiency. Further, language of instruction has little meaning in the out-of-school population.

Observation 3: Extensive consultation and participant involvement in test development activities have been at the core of PISA. The extent of consultation with potential developing country participants and their capacity to influence PISA design choices needs to be given careful consideration.

Observation 4: The item-by-country interactions (country DIF) appear to be enormous between developing countries. This has severe implications for the validity of described scales and for construct comparability more generally. In the field trial the potential impact of these interactions on the validity of reporting should be examined and if appropriate alternative reporting schemes be considered.

Observation 5: The PISA tests are set at quite a high difficulty level, relative to typical student performance. In the case of countries that perform less well the average percent correct on the items is very low and assessing student with such a test is clearly inappropriate.

Observation 6: The pool of secure PISA items is well targeted in terms of optimising the average measurement precision across all participants.

Observation 7: The available secure item pool has an information profile that does not match the likely proficiency profile in candidate PISA for development countries. It follows that utilising a test design that results in administering each of the existing secure items to an equal number of students would not be efficient.

Observation 8: Drawing upon easy items only it appears that test designers will face challenges in building a test that matches the framework specifications. The implications in terms of preparing an assessment that is fit for purpose may not be profound, but it does suggest that it will not be possible to report at the subscale level.

Observation 9: If an easy subset of items that approximates the framework is selected from the secure pool it will remain more difficult than is psychometrically ideal for many developing countries – *i.e.* with the smallest possible measurement error. In other words the test will be mis-targeted.

Observation 10: A rotation scheme with fewer booklets than is used in regular PISA will be sufficient due to the limited pool of suitable testing material.

Observation 11: Without access to computer-delivery and working within the constraints of the available secure item pool a simplified test rotation design could be developed to provide domain level, but not sub-domain level proficiency estimates.

Observation 12: The limited size of the available secure pool will add instability to trends over time and the comparability of the results across countries. A possible way to increase the size of the item pool will be to use good easier items that have been rejected in previous PISA surveys after the field trial. They may have had good item characteristics but were rejected for other reasons - perhaps to ensure an adequate balance of item difficulty or because the framework was sufficiently covered by other items. No new development would be needed for these items.

Observation 13: The requirement of testing 15-year-olds who are not in school has major implications for the test design. It is likely that a separate test will be designed for these students. The possibility of delivering the test by computer can also be considered using a model similar to that used in the PIAAC project. In addition the level of the test will need to be carefully considered. The PISA Reading Components assessment may be more appropriate to administer to out-of-school students. This has the disadvantage of testing reading only and lacking a real link to PISA because only two countries have participated in this assessment.

Observation 14: It will be essential to implement an equating study between the PISA instruments and the PISA for development instruments so as to document the comparability.

Observation 15: The current PISA described proficiency levels in reading do not provide enough useful information for many developing countries - making it difficult for policy makers to identify and implement remedial interventions focused on their students' weaknesses.

Observation 16: In mathematics, in some countries, nearly half the students are below the lowest level for which PISA can describe student capacity.

Observation 17: When comparing reading, mathematics and science it is the last two which have the largest percentage of students below a described proficiency level - this is partly due to the fact that the described level 1 for reading was extended and divided into two sub-levels.

Observation 18: The fit of the developing country data to the PISA model is not good and scaling modifications to address some of the deviations should be explored. Changes to the model would however have wider implications for PISA, including a need to rescale previously collected data.

Observation 19: The use of learning metrics to describe dimensions of educational progression is at the core of the PISA reporting methodology. This approach to reporting and construct validation requires a consistency across countries in item behaviour than is not apparent for developing countries.

INTRODUCTION

Background

The OECD is seeking to enhance its Programme for International Student Assessment (PISA) to make it more relevant for contexts found in developing countries. Since the inception of PISA in 2000, there have been an increasing number of countries participating in each survey – these extra countries have been, almost exclusively, non-OECD countries.

Table 1. Participants in each PISA assessment

Year	Number of OECD Countries	Number of non-OECD countries
2000	28	15
2003	30	11
2006	30	28
2009	34	30
2012	34	31

PISA currently builds on the experience of over 70 countries, among them several low- and middle-income countries. The original framework for PISA was based on OECD countries, but as the membership in PISA expanded, the frameworks evolved as well. All countries that take part in a PISA assessment also participate in the framework development. However, with an increasingly diverse group of countries, the targeting of the assessment may be less appropriate for lower-middle-income countries and low-income countries.

With the PISA for Development project, survey instruments will be developed that are better adapted to developing country contexts. These adaptations will better enable more developing countries to accurately assess and describe 15-year-olds' competencies, knowledge and skills in the key subjects of reading, mathematics and science, while at the same time providing the countries with an opportunity to build their capacity to manage student assessment and apply the results for system improvement. One of the main challenges will be to construct enhanced and adapted cognitive instruments in reading, mathematics and science, while still obtaining robust results that are comparable to the international PISA scales.

The purpose of this paper is to identify the main technical issues to be addressed with regards to enhancing the descriptive power of PISA's cognitive assessments in reading, mathematics and science, particularly considering the contexts (and students) likely to be encountered in developing countries. This includes, but is not limited to, considering the higher percentage of students that may perform below baseline proficiency levels in PISA, the capacity of PISA to describe the different levels of performance of these students, and to adequately include in the assessment students (*e.g.* during the data collection phase) at the lowest ends of performance.

The report consists of five sections. The first section provides an overview of the development of assessment frameworks and items and describes the related quality assurance procedures. It concludes with some comments on the cross-cultural validity of PISA. The second section provides a review of the available pool of secure PISA items in reading, mathematics, and science. It reviews the appropriateness of the difficulty of those items and the possibility of covering the existing PISA frameworks with a subset of the items that has appropriate difficulty for use in developing contexts. In the third section we provide an overview of the current test design and raise some issues concerning adaptations to the design for developing contexts. In the fourth section we review the appropriateness of the proficiency levels that are the core of the PISA reporting, in terms of their utility for developing countries. Finally, in the fifth section we discuss PISA scaling models and their applicability in developing contexts. Most observations in the paper are derived from the authors' experience with, particularly PISA 2012, but also from their involvement in all preceding PISA surveys.

Country experience in international assessments

International surveys have been part of the education research scene since the 1960s. Table 2 summarises the major characteristics of the international educational assessments.

Table 2. Major characteristics of international educational assessments

Assessment	Organiser	Subjects assessed	Target population	Years of administration
EGRA	WORLD BANK	Basic literacy skills	Early years	Continuous
ICCS	IEA	Civic knowledge and attitudes	Grade 8	2009
PASEC	CONFEMEN	French, mathematics	Grades 2 and 5	
PIRLS	IEA	Reading	Grade 4	2006, 2011
PISA	OECD	Reading, mathematics, science, problem solving	Students aged 15	2000, 2003, 2006, 2009, 2012
SACMEQ	SACMEQ Assembly	Reading, mathematics	Grade 6	1995, 1998, 2008
SERCE	LLECE	Reading, mathematics, science, writing	Grades 3 and 6	1995, 1998, 2008
TIMSS	IEA	Mathematics, science	Grades 4 and 8	1995, 1999, 2003, 2007, 2011

The countries which have expressed an interest in participating in the PISA for Development project, Cambodia, Ecuador, Guatemala, Senegal, Sri Lanka and Zambia, all conduct assessments for their students at either local or national level. Some of them also have experience in international assessments. It is this experience which should provide a suitable foundation for participation in this project. The countries' experience is shown in Table 3 with major assessments listed. While most countries have undertaken at least one international assessment it will be a requirement that a large degree of capacity building will be incorporated into the PISA for Development project.

Table 3. PISA for Development countries' experience in international educational assessments

Country	EGRA	ICCS	PASEC	PIRLS	PISA	SACMEQ	SERCE	TIMSS
Cambodia			●					
Ecuador							●	
Guatemala							●	
Senegal	●		●					
Sri Lanka								
Zambia						●		

DEVELOPMENT OF ASSESSMENT FRAMEWORKS AND ITEMS

The PISA for Development assessment will be carried out in the three domains of reading, mathematics and science with equal weighting given to each domain. For the initial part of the project there will be no new items developed. The established assessment frameworks therefore provide the foundation of the tests and an understanding of their development is important.

Assessment frameworks

The development of all PISA assessment items is based on assessment frameworks for each of the subject areas. Assessment frameworks have been developed not only for reading, mathematics and science, but also for problem solving and financial literacy.

An assessment framework is a statement and discussion about what an assessment intends to measure. From the test development perspective, a framework is particularly important because it provides an initial guide for task development and at any later point can be used as a reference for evaluating the extent to which the instrument has met its purpose. Typically the development of a subject area assessment framework is guided by a group of internationally recognised experts. The work of the expert groups is coordinated by the international contractor and submitted to the PISA Governing Board for approval. Arriving at a final statement for the assessment framework, therefore, is a collaborative process that involves the participating countries. Framework development for a large-scale educational assessment is best led by an expert group comprising members from a range of backgrounds including those with academic subject expertise, experience in large-scale assessment development and subject-area pedagogical knowledge. The composition of the subject expert groups is extremely important because it has a direct influence on the material that will be included for assessment. Membership of the groups is determined firstly on the basis of expertise in the area of assessment, so that the resulting expert group will be made up of experts who are internationally recognised leaders. At the same time, without sacrificing the level of expertise, there is a desire to ensure that the experts are not from just one or two countries or from just one or two language backgrounds.

At the same time, it is important that test developers are included in the expert group, or at least attend expert group meetings, so that they gain understanding of the theory underlying the framework and can provide advice about what is practical and feasible

Frameworks normally start with a definition of the assessable domain, followed by an elaboration of the terms of the domain, and then an outline of the major variables to be included.

Framework development should occur in conjunction with test development. While the framework may be initiated before test development begins, to provide some structure and guidance at an early stage, it is important that the process includes the capacity to review and revise the framework in light of its application, as the instrument is developed.

In recent PISA surveys there has been a trend to increase the separation between framework development and item development. This occurred in PISA 2012 when both the item development contractor, ACER, worked in conjunction with another organisation not involved in item development - the Achieve Corporation. In 2015, development of the frameworks was a separate module and was handled by the Pearson Corporation. At the time of writing it is believed that framework development will continue as a separate module in PISA 2018.

> Observation 1: In any move to expand the use of PISA to a greater number of countries it would be essential to carry out a complete review of the assessment frameworks in consultation with those countries. It might be expected that the areas currently included for assessment which are seen as priorities by OECD countries may not coincide with the areas that are seen as priorities for developing countries. At the same time any extension of the framework will need to continue to incorporate the original.

While the document which will be supplied to the participating countries, “Overview of the Project Implementation Plan”, describes the process through which participating countries will be consulted, it does not specifically refer to the PISA Assessment Frameworks. Countries should, while planning their future analysis and reporting, consider the relevance of the areas described in the assessment frameworks. Assuming that the test which is administered to the students may be a subset of the existing PISA item pool, feedback from countries on relevance of different parts of the assessment frameworks will guide those who are composing the tests.

Country involvement in this process will also go towards the capacity-building approaches in this project - the development of a Capacity-Needs Analysis and a Capacity-Building Programme as outlined in the document, “Overview of the Project Implementation Plan”.

Most of the analysis presented in this paper is based on the most recent PISA survey for which data are available - PISA 2012. The PISA assessment framework for PISA 2012 is publicly available in the document *PISA 2012 Assessment and Analytical Framework Mathematics, Reading, Science, Problem Solving and Financial Literacy* (OECD, 2013). For PISA 2015, computer-based assessment will be the primary mode of delivery for all domains. However, paper-based assessment instruments will be provided for countries choosing not to test their students by computer. The Reading Literacy component for both the computer-based and paper-based instruments will comprise the same intact clusters of reading trend items. With the move to computer-based delivery for 2015, the 2009 Reading Literacy text classification ‘medium: print and electronic’ has been updated to ‘fixed-text’ and ‘dynamic-text’ to distinguish between delivery mode and the space in which the text is displayed, regardless of whether it is printed or onscreen. It is important to note, however, that the constructs of the Reading Literacy Framework remain unchanged. For PISA 2015, new Scientific Literacy items will only be available in the computer-based assessment. However a paper-based assessment instrument will be provided for countries choosing not to test their students by computer which will consist only of the trend items. The mathematical literacy component for both the computer-based and paper-based instruments will comprise of the same intact clusters of mathematics trend items.

While the PISA for Development assessment will be a paper based assessment it will be necessary to consider the implications of any modifications to the assessment frameworks.

PISA item development

Once a draft framework has been prepared test development begins by assembling a broad range of tasks that address explicit or implicit understandings of what the domain is all about (Mendelovits, forthcoming).

Part of the test developer’s early work is to accompany every drafted task with an explicit statement, related to the framework, about what the task aims to measure. A particularly important aspect of test development for an international assessment administered in several languages is knowledge about translation. Tasks need to be framed in such a way as to facilitate equivalence across languages.

The salient issues for test development in large-scale assessments can be summarised as:

- Ensuring construct validity: that the test is measuring what it purports to measure.

- Ensuring that the instrument is fair to all test takers, so that no disadvantage or advantage arises because of national, cultural, linguistic, gender or socio-economic variation.

- Balancing the claims of various stakeholders.

- Balancing the demand of reporting trend with the capacity to innovate.

One of the strong aspects of PISA item development has been the involvement of the participating countries. This occurs in three main ways.

Firstly, the countries are invited to submit ideas for questions to be included in the assessment. This gives the countries the opportunity to put forward questions which are of particular interest to them or reflect their curriculum. In this context it should be remembered that PISA is not a curriculum based test - the assessment framework is generated by a consultative process not a process of finding common curriculum components.

Secondly, all countries are involved in the assessment of the items through a process of item review. This review process ensures that every country has a say on the appropriateness of the items for their students.

Thirdly, all countries have the opportunity to contribute to the discussion on coding of the items at international training meetings. These meetings are also especially useful to the item developers who get the opportunity to hear about unforeseen difficulties with items.

There are three main stages in development of the items before they are used in a field trial:

1. *Panelling* – where fellow test developers review and critique each other’s items from every perspective they can think of.

2. *Cognitive interviews* are sessions conducted by test developers in which individual students, or very small groups of students, as similar as possible to the test’s target population, are administered a few draft assessment tasks, and asked to talk about their thought processes, either as they do the tasks or immediately afterwards. Test developers sit with the students and record their responses.

3. *Pilot testing* is a second way of gathering information about test taker responses to refine tasks at a relatively early stage. In this procedure test material is administered to larger groups – perhaps a regular class size. With class-sized groups of 20 to 40 students the sample is not big enough for fine-grained statistical analysis of student response, but the process does start to give some idea of the relative difficulty of various tasks.

PISA field trialling

Field trialling takes place with two main aims in mind:

1. *Assessing the suitability of the items that have been developed* – both classical analysis and Item Response Theory (IRT) analysis provide data that are used to investigate the quality of the tasks. Using the output of these tools, test developers inspect the fit and discrimination of each task, to ascertain whether the task is measuring something similar to other tasks in the assessment – that is, whether it is contributing to the measurement of a coherent underlying trait (such as scientific literacy or problem-solving ability). If the trial test sample is of a sufficient size, it is also possible to analyse the data to ascertain whether there is differential item functioning (DIF) among identified subgroups. It is important to note that DIF analysis does not focus on the overall differences in performance between subgroups, but rather – given any underlying overall difference – on items or tasks that have elicited anomalous performance. In each country a sample of around 1000 students is selected to participate. This number is chosen because it will give sufficiently reliable statistics on each new item's characteristics. The field trial, like the main survey uses a rotated booklet test design so that not all students do all items, but a sufficient number of students encounter each new item.

2. *Testing the capacity of a PISA national centre* to carry out the logistics of the implementation of the assessment. Implementing a large scale educational assessment such as PISA is an exercise that needs the timely deployment of many resources - both physical and human. For countries not familiar with this activity the field trial provides a means to gain this experience and to modify procedures before the implementation of the main survey.

In large-scale assessments, up to three times as much test material is trial tested as the amount needed for the main administration, to allow for the attrition of items that have shown poor performance in statistical terms, and, given that, to ensure that enough sound material remains to fulfil all the framework requirements with the final selection of tasks.

PISA final item selection

In PISA approximately one half of the items that have been field trialled will be discarded in the final selection process. The process of selection of the items for the main survey is a process that needs to keep in mind many constraints. Currently in PISA there are 13 booklets each with four clusters of assessment material. Booklets can be linked through having clusters repeated in more than one booklet.

Each cluster needs to be of equivalent difficulty and should take the same amount of time for students to complete - in PISA clusters take 15 minutes. At the same time across the range of clusters (there are seven clusters for the subject domain which is the major focus of assessment – reading in 2009, mathematics in 2012 and science in 2015) all components of the framework must be covered in the proportions agreed to by the PISA Governing Board. On top of this there needs to be a mix of multiple-choice and constructed response items.

When all of these constraints are met the clusters need to be arranged through the booklets in a manner which allow results of students to be linked.

PISA language of assessment

Generally in PISA the language of assessment is the same as the language of instruction. This only varies where there is ambiguity or variability in the language of instruction.

Children come to school with diverse language backgrounds. In some countries children may arrive at school with limited, or perhaps no, experience of the language of instruction because the language spoken at home or in the local community is different. The consequence is that the proficiency of these students may be less validly assessed. Generally students are excluded from the assessment if they have not had one year's experience at school of the language of the assessment.

In some countries different languages are used in different subjects. This issue has been handled in different ways - in some cases students have been asked to nominate a single language in which they should be assessed - this can put them at a disadvantage in the subjects where they are less familiar with the language. In some cases countries have chosen to print hybrid booklets which include different languages matched to the language of instruction. This can be a complex administrative task when assembling the booklets, but achieves the goal of language of assessment being the language of instruction.

For students who are not in school, the term 'language of instruction' has little meaning and the assessment language will need to be the language spoken at home.

> Observation 2: The PISA standard of testing is the language of instruction should be reviewed for PISA for Development since there may be large number of students for which an assessment in the official language of instruction does not provide a valid indication of a student's proficiency. Further, language of instruction has little meaning in the out-of-school population.

PISA translation and adaptation

When the item set for the field trial has been decided it is necessary to translate each item into French so that there are two source versions of every item – English and French. This process has been found to be extremely helpful in the final preparation of the item as sometimes language difficulties remain unseen in one language until they are translated into another.

In PISA a process of double translation is undertaken to ensure the comparability of items across the many languages in which PISA is administered. The first step in this process is for the participating country to translate into its own language a version of the test from the English source version and from the French source version. These two versions are then compared to arrive at a final version. This final version is then verified by an international language expert consultant to ensure that the version of the items being used in the test truly reflect the nature of item and will not advantage nor disadvantage the students undertaking the test.

Following the field trial the item analysis may reveal unusual levels of difficulty (high and low) which can often be traced back to an inappropriate translation.

It is also necessary to adapt some components of the questions to overcome differences in word use and culture (for example some English-speaking countries use the expression “mobile phone” while others use “cell phone”). Adaptations are also necessary in questionnaires to account for different education programmes and systems, and word usage (this is covered in the questionnaire strand).

In PISA the above described procedures are supported by a battery of trial data analyses implemented to ensure the cross-cultural validity of the assessment. The components of the process include:

- The use of international experts to guide framework and item development.

- Extensive framework and item review opportunities by all participants.

- Submissions of items actively sort from all participants with high priority given to the use of participant submissions.
- Engagement of professional test development teams from many countries.
- A requirement that all items are trialled by all participating economies.
- The implementation of extensive linguistic adaptation and verification.
- Careful psychometric review of all items.
- Examination of item-by-country interactions in both Field Trial and Main Survey.

Observation 3: Extensive consultation and participant involvement in test development activities have been at the core of PISA. The extent of consultation with potential developing country participants and their capacity to influence PISA design choices needs to be given careful consideration.

For the PISA for Development project, each participating country will draw up, in consultation with the OECD and appointed contractors, an implementation programme that should include review of item suitability for each country, how translations will be conducted and verified. There will be some differences to the normal PISA process. It is possible that there will be some variation in the implementation timeline as opportunities arise in each country for the best testing window.

Cross-cultural validity

Despite the repeated expression of the importance of the above described quality assurance steps input from countries is uneven and it is dominated by western countries and Indo-European languages. The issue of cross cultural differences is a significant challenge to any international assessment programme.

In PISA all participating countries are given the opportunity to review the items which are being proposed for the assessment. In addition to reviewing the items for difficulty and curriculum appropriateness, the countries consider whether they are culturally acceptable or not for their student body. This process is extremely important because the developers of the original items may not be aware of particular sensitivities that may exist in some countries.

In addition, the process of designing and implementing PISA is a collaborative one, where, at the operational level, all countries are given the opportunity to provide specific input regarding their concerns with the test or the contextual questionnaire. The PISA for Development project will follow this collaborative process.

The most systematic report on the issues of the cross cultural validity of PISA was undertaken by Grisay *et al*. (2007). They show that the combination of a Non Indo-European test language and a large proportion of low-achieving students in the country influence the pattern of item difficulties in PISA tests and, by doing this, reduce comparability between such countries and the main body of PISA countries.

In particular, Grisay *et al*. (2007) used PISA 2006 field trial science data to investigate the amount of difference in the pattern of item difficulties between participating countries and possible reasons for such differences. To quantify the amount of difference in the pattern of item difficulties between participating countries the authors defined a uniqueness indicator for each participating country as 100% minus communality found from the principal component analysis, minus completely random error found from simulation. The uniqueness indicator was then regressed onto a number of factors:

- *Language of test grouped into two categories: Indo-European and other*; language of test was chosen as a proxy for linguistic and cultural differences; two categories were left because the use of more specific categories such as Germanic, Romance, Slavic Altaic and Finno-Ugrian did not improve the correlation between the language of test variable and the uniqueness variable.

- *The country's GDP per capita expressed in US dollars* was chosen as a proxy for possible economic differences; if a country tested students in more than one language, the same GDP per capita was used for each national version.

- *Number of key corrections required by verifiers per 100 000 characters of text*; most of corrections were implemented before the field trial but "the hypothesis behind the indicator was, however, that those national versions were large numbers of serious translation errors had been identified and corrected were more likely than other versions to contain residual translation errors".

- *Curriculum coverage* was defined as average rating across 247 items for each country; national project managers as a part of the national review of the field trial items rated each item from 1 (not included in the national curriculum for 15-year-old students) to 5 (perfectly appropriate for 15-year-old students in the country).

- *Average item discrimination* was used as a proxy for possible targeting effects.

The results showed that the *Uniqueness* variance can be decomposed approximately as follows: "*Non Indo-European language* [of test], 24%; *Average Item Discrimination*, 10%; *Curriculum Coverage* and *Key Corrections*, 2%; variance explained jointly by *Non Indo-European language* and *Average Item Discrimination* (also partly confounded with GDP, Key Corrections and Curriculum Coverage), 35% and, non-explained variance: 29%".

In other words, nearly 70% of uniqueness of the national item difficulty pattern is explained by Non Indo-European language of test and average item discrimination. The latter in turn can be explained by large proportion of low achieving students. When analysis was repeated using 75% of easiest items the uniqueness of the pattern of item difficulties increased.

An illustration of these differences is seen in Figure 1 where the percent correct on all PISA items used in 2009 is compared for a selection of countries. In this figure we are using a percent correct, rather than a logit or PISA scale score metric. This was done for ease of communication and if alternative metrics where used the central observations would be identical.

Figure 1 (i) shows the relationship between percent correct in the United States and the OECD Average. The plot shows very little spread around the identity line (shown in blue) and the correlation between the values is 0.961. The implication is that there is consistency in the construct definition between the US and the average of the OECD. Figure 1 (i) also shows that the majority, but far from all, of the items were slightly harder in the US than for the OECD on average.

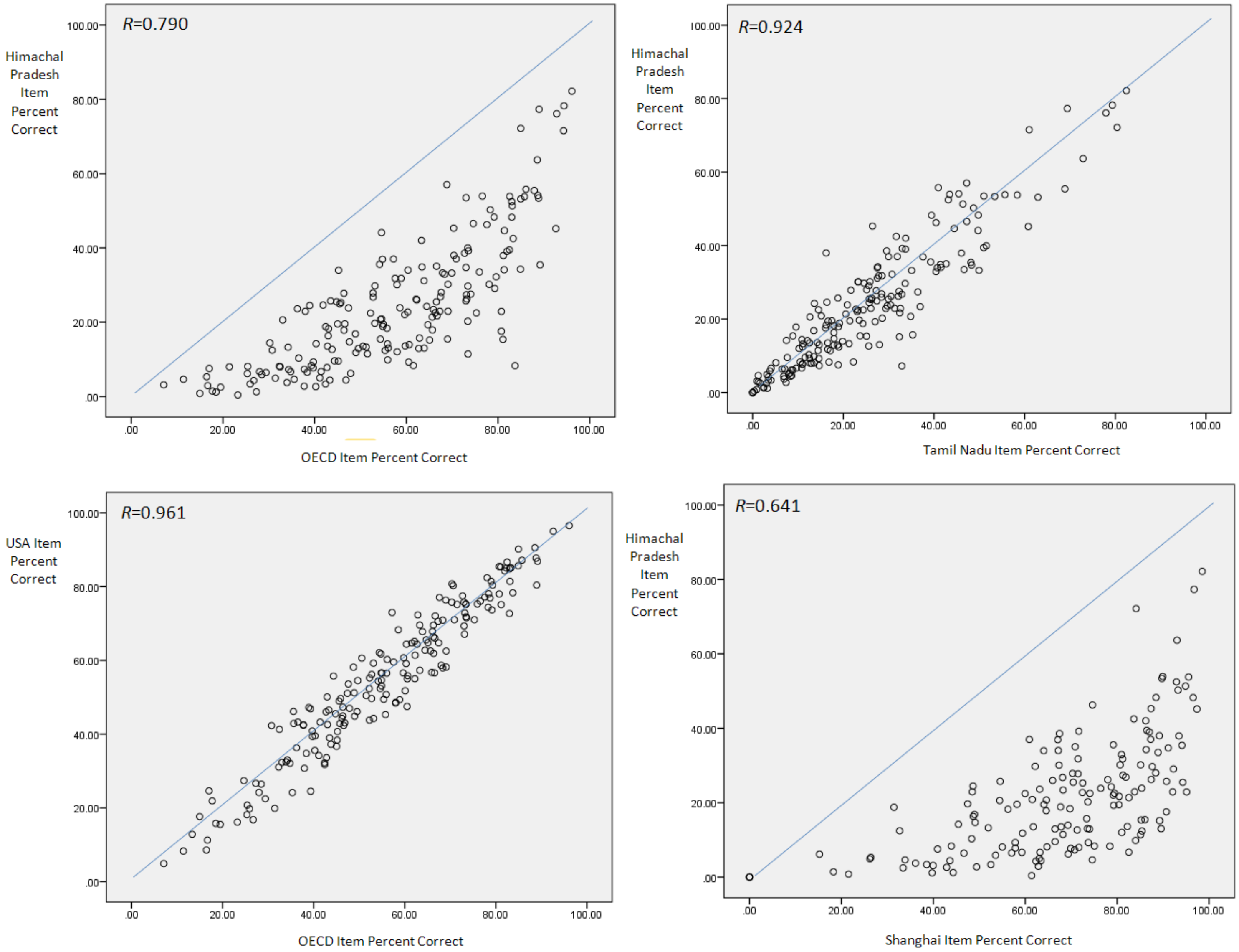
Figure 1 (ii) shows the relationship between percent correct for the Himachal-Pradesh state in India and the OECD Average. The plot shows that every item was harder in Himachal-Pradesh than for the OECD on average and for many of the items the difference is considerable. The correlation between the values at 0.790 is much lower than for the US, supporting the concerns of Grisay *et al*. (2007) regarding the stability of the PISA constructs across such markedly different cultural settings as OECD countries and Himachal Pradesh.

Figure 1 (iii) shows the relationship between percent correct for the two Indian states that participated in PISA 2009 – Himachal Pradesh and Tamil Nadu. The correlation between the values, at 0.924 is quite high, showing construct consistency within India, but across languages. Finally, Figure 1 (iv) shows the relationship between the percent correct in Himachal Pradesh and Shanghai (China). The figure highlights both the difference in performance level and the nature of the construct.

The data discussed above relates to countries which have already participated in PISA. As part of the process of implementing the PISA for Development pilot project a field trial will take place. This field trial will provide information with respect to the participating countries' interactions with items. The data from this trial should be used to quantify the potential impact of item-by-country interactions on validity of PISA for development reporting. This could be done using procedures such as those employed in Adams, Berezner and Jakubowski (2010). If the threats to validity are found to be unacceptable then it may be necessary reconsider the current approach to reporting PISA results and retreat to a basket of goods (or market basket) approach (Mazzeo, Kulick, Tay-Lim, Perie, 2006).

> Observation 4: The item-by-country interactions (country DIF) appear to be enormous between developing countries. This has severe implications for the validity of described scales and for construct comparability more generally. In the field trial the potential impact of these interactions on the validity of reporting should be examined and if appropriate alternative reporting schemes be considered.

Figure 1. Comparison of the percent correct on PISA 2009 items for a selection of participants



Review of secure item pool[1]

Over the administrations of PISA in 2000, 2003, 2006, 2009 and 2012 a total of 517 items have been used in the estimation of the proficiency distributions. Of those items 180 items have been released into the public domain, leaving 337 secure items for potential future use.

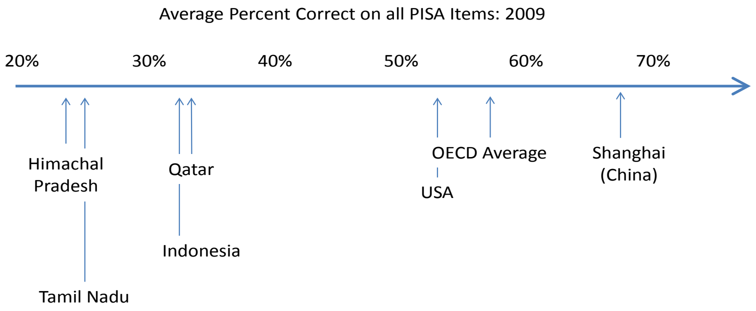
Table 4. Number of unique items used in PISA assessments, number of items released and number of items secure

	Number of different items used	Number of released items	Number of secure items
Reading	223	80	143
Mathematics	169	64	105
Science	125	36	89
Total	517	180	337

Later we review the extent to which the pool of available secure items covers the PISA frameworks. We first here examine the characteristics of the pool in terms of difficulty.

The evidence reported in the PISA technical reports (Adams and Wu, 2002; OECD, 2005; OECD, 2008; OECD, 2011) shows that the PISA tests have, on average across all countries, been quite challenging for students. This has been caused by the mismatch between the expectations and aspirations of leading educators and the actual performance levels of students. In the case of lower performing countries the number of items that students can typically respond to correctly is quite small, as is illustrated in Figure 1. This figure shows the average percent correct for students in a small selection of countries for PISA 2009. The average reported in the figure is for all items. The major domain in PISA 2009 was reading and reading has typically had the highest percent correct values, the implications is that the values would be lower than those shown here for years in which mathematics and science are major domains.

Figure 2. Illustration of the difficulty of PISA tests for a selection of participants



[1] It is important to note that the ongoing security of PISA items is not a trivial matter and that procedures must be put in place which guarantee that any access to the secure items will not in any way threaten the implementation of the normal three-yearly PISA surveys.

> Observation 5: The PISA tests are set at quite a high difficulty level, relative to typical student performance. In the case of countries that perform less well the average percent correct on the items is very low and assessing student with such a test is clearly inappropriate.

Given that there is no new test item development, the challenge for the PISA for Development project will be to use the existing items in such a way to better assess those students whose capacity is at the lower end of the scale.

No changes to the stimulus of items will be allowed as this will have the effect of creating totally new items which will not be comparable to the original PISA items. Experience shows that the slightest change to item stimulus (or even layout) affects the characteristics of an item.

Information targeting

A useful way to illustrate the appropriateness of the match between the item pool and the student proficiency distributions is to estimate the item pool information function, using all items that have been used in PISA. The item pool information function describes how accurately the item pool is able to distinguish between students over the range of the proficiency distribution.

The estimation of the item pool information function requires that the location of all items be estimated on the PISA scale. To achieve this it was necessary to undertake an item response theory scaling of all PISA items. This scaling was undertaken using ACER ConQuest (Adams, Wu and Wilson, 2012) following the procedures that have been implemented in each PISA assessment to date and using a pooling of the calibration samples from each PISA assessment. This rescaling was necessary because the equating methodologies used for PISA do not provide a single set of item locations for all items.

After placing the items on a common scale the information function was compared to the proficiency distribution.

Figure 3 shows this comparison for reading using only the secure items. The left panel of the figure shows the estimated reading proficiency distribution based upon the pooled PISA calibration samples. The right panel shows the information function for the pool of secure reading items. The information function for all items is not shown, but is essentially identical to that for the secure items.

The figure shows that the peak in the reading information function is below the mean of the proficiency distribution. That is the item pool is at its most informative for students with proficiency estimates that are just below the mean of the calibration sample.

Figure 3. Comparison of the item pool information function and the calibration sample proficiency distribution for reading

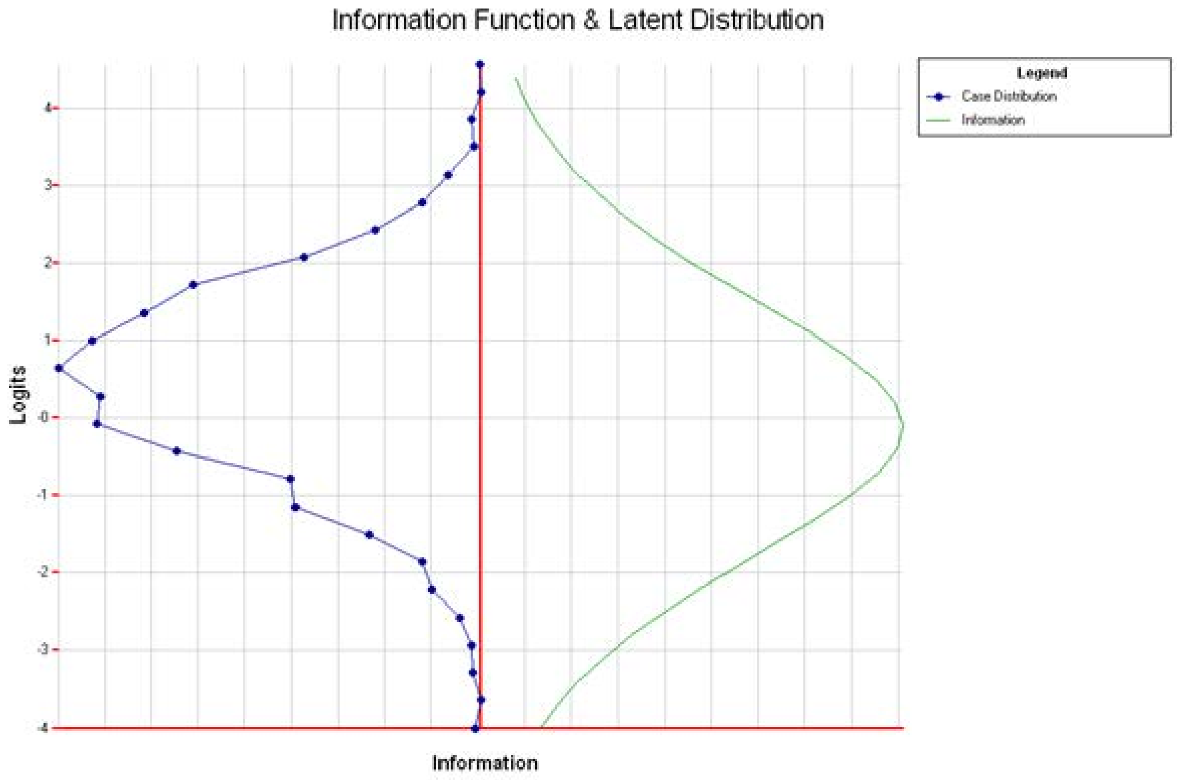


Figure 4 shows the same comparison for mathematics, again, using only the secure items. In this case we note that the peak in the information function is above the mean of the proficiency distribution. That is the PISA mathematics pools is at its most informative for students with proficiency estimates that are just above the mean of the calibration sample.

Figure 5 shows the case of science, where the relationship is similar to that for reading.

Figure 4. Comparison of the item pool information functions and the calibration sample proficiency distribution of mathematics

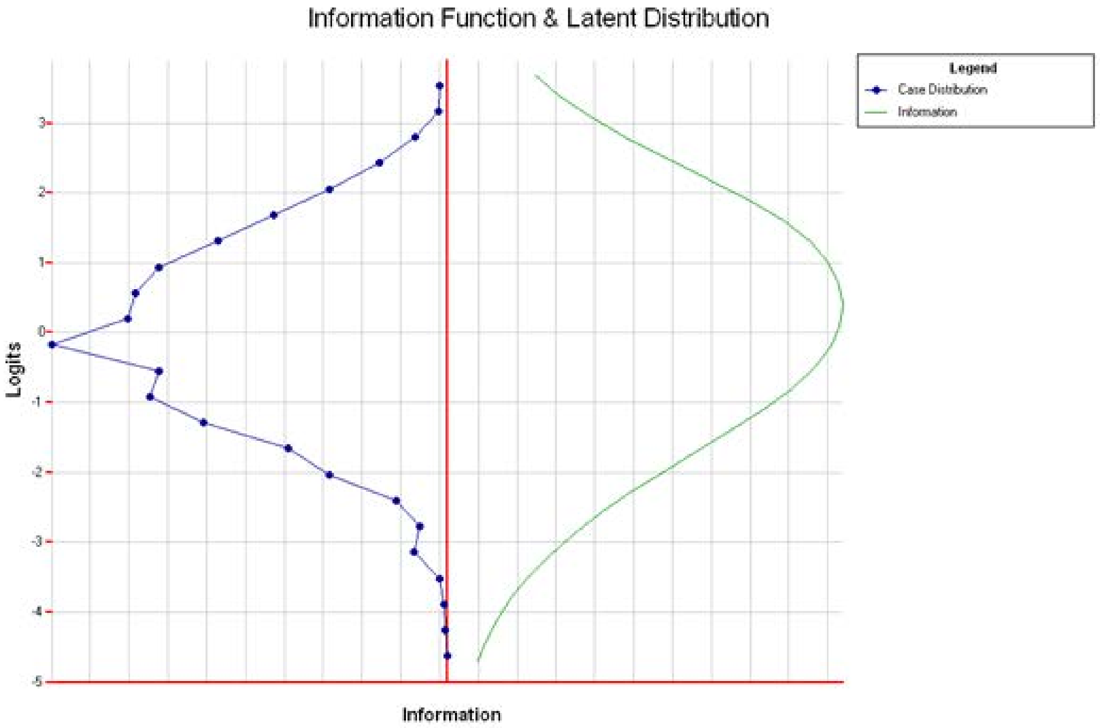
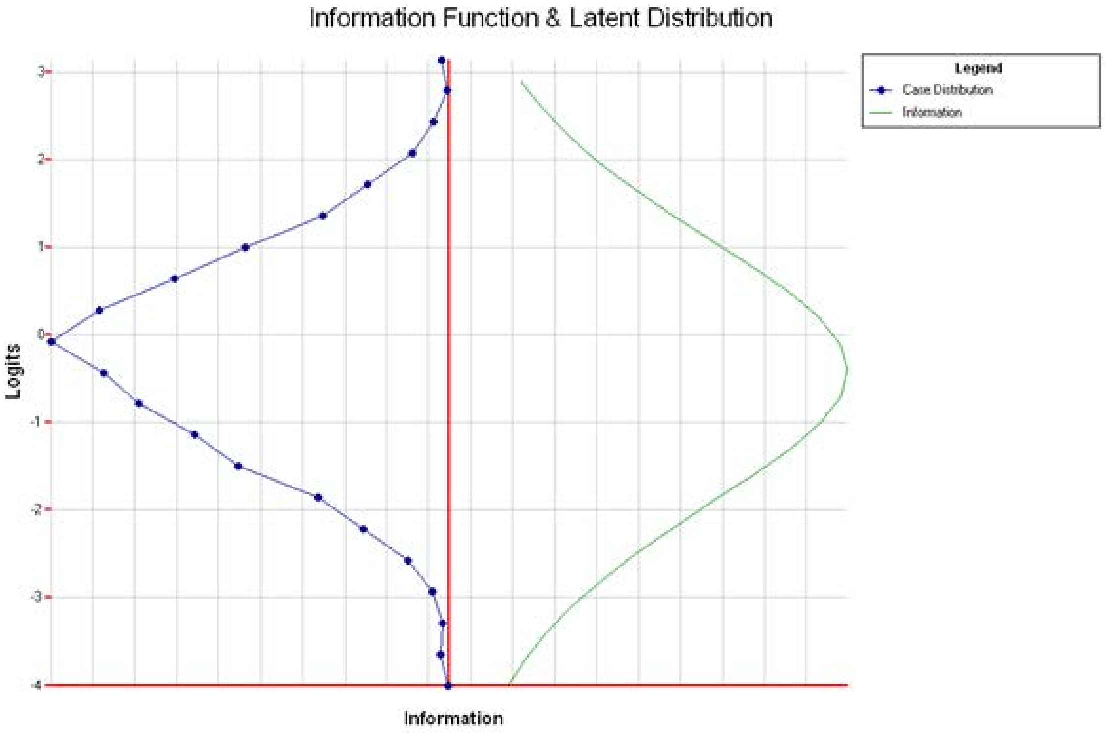


Figure 5. Comparison of the item pool information functions and the calibration: Sample proficiency distribution for science



> Observation 6: The pool of secure PISA items is well targeted in terms of optimising the average measurement precision across all participants

In the context of PISA for development, however, our concern is the relationship between the information function and the proficiency distribution in countries that do not perform highly on PISA. This relationship is illustrated in Figure 6 which shows the relationship between the information functions and the estimated proficiency distribution for Kyrgyzstan in 2009.

We have selected Kyrgyzstan 2009 for illustrative purposes and do not have any firm evidence concerning how well the countries that are proposing participation in PISA for Development might perform. In PISA there are examples of high-income countries that perform quite poorly – Qatar, and examples of lower-income countries that perform well – Viet Nam. The field trial will be crucial in determining the likely performance levels of PISA for Development countries and will be quite influential in finalising the test design.

In this case we see a marked discrepancy with the peak of the information well above the proficiency distribution for all three domains. To illustrate the discrepancy, Table 3 shows the proportion of total secure mathematics pool information available to each tenth of the Kyrgyzstan proficiency distribution. Each interval shown in the table contains 10% of the student proficiency distribution, yet with exception of the first tenth and the last tenth the amount of information is well less than 10%. In fact, the PISA test pool has 43% of its information available above the 90th percentile of the Kyrgyzstan distribution. The median of the distribution is –1.27, so the 32% of the information is available for the lower half of the distribution and 68% for the upper half

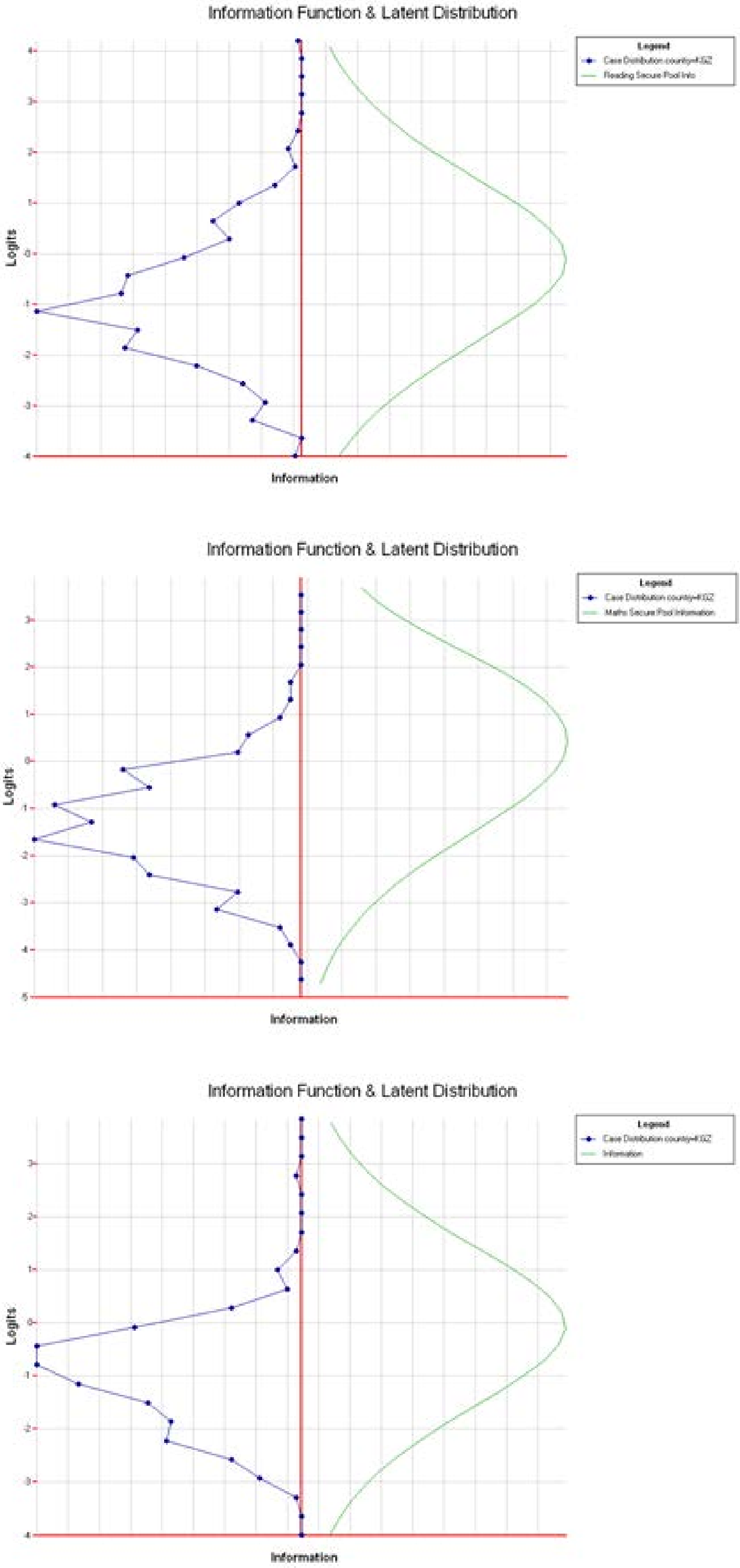
Table 5. Proportion of total information available provided by secure mathematics pool for each 10th of the Kyrgyzstan mathematics proficiency

	Interval									
	Less than -2.55	-2.55 to -2.12	-2.12 to -1.91	-1.91 to -1.59	-1.59 to -1.27	1.27 to -1.06	1.06 to -0.74	-0.74 to -0.42	-0.42 to -0.11	Greater than -0.11
Proportion of information	0.14	0.05	0.03	0.05	0.05	0.04	0.08	0.07	0.06	0.43

Clearly the item pool is not optimal, in terms of difficulty for use in Kyrgyzstan. The average percent correct across all items on 2009 in Kyrgyzstan was 27%, which is slightly higher than for the two Indian states shown in Figure 2.

> Observation 7: The available secure item pool has an information profile that does not match the likely proficiency profile in candidate PISA for development countries. It follows that utilising a test design that results in administering each of the existing secure items to an equal number of students would not be efficient.

Figure 6. Comparison of the item pool information functions (for reading, mathematics and science) and the calibration sample proficiency distribution for Kyrgyzstan



Framework coverage of secure items

If the 337 secure PISA items available from 2000 to 2012 are classified according to the most recent frameworks we find that in most regards the pool provides the same coverage as the whole pool. The concern, however, is that the pool is not well targeted to low-performing countries. In this section we consider the implications, in terms of framework coverage of restricting attention to subsets of items that would be well targeted at low-performing countries.

The distribution of the secure items according to PISA level is shown in Table 6.

Table 6. Distribution of secure items by difficulty level

PISA level	Proportion of reading secure items	Proportion of mathematics secure items	Proportion of science secure items
6	4%	19%	3%
5	9%	14%	6%
4	22%	21%	34%
3	49%	21%	36%
2	41%	17%	17%
1	NA	8%	4%
1a	6%	NA	NA
1b	20%	NA	NA

Coverage provided by selections of targeted items

For the sake of this section we have assumed that the goal was to form a single two-hour booklet which would contain about 60-65 items, one third from each of Reading, Mathematics and Science. The goal therefore is to find a set of items that would cover the framework, yet be well targeted to lower performing countries. In the interest of making the selection as simple as possible we have focussed on coverage of the reporting aspects of the respective frameworks.

In tables 7-11 the distribution of items over the reporting components of the frameworks is shown for selections of easy items and this is compared to the entire secure pool and the original framework intentions. We have chosen the least difficult 37 reading items, 32 mathematics items, and 45 science items. In the case of reading we have chosen two sets because the restricted set provided a clearly inadequate coverage of reflect and evaluate. The second, and larger set, provides a modest improvement. These items largely come from levels one, two and three.

It is important to keep in mind that the tables presented here represent a best case scenario. In selecting easier sets we have relied upon item difficulty only, we have not taken into the account the unit structure, nor have we considered the constraints that might be imposed through issues of cultural appropriateness.

Table 7. Number of secure reading items broken down by the aspect sub-scales

	Number of secure items	Proportion of total	Number of relatively *easy* (1) secure items	Proportion of total for relatively *easy* (1) secure items	Number of relatively *easy* (2) secure items	Proportion of total for relatively *easy* (2) secure items	Target proportion in most recent framework
Access and retrieve	42	29	19	51%	29	40	22%
Integrate and interpret	71	50	16	43%	35	48	56%
Reflect and evaluate	30	21	2	5%	9	12	22%
Total	143		37		73		

Within the reading pool the *access and retrieve* items are typically the least difficult, followed by *integrate and interpret* and then *reflect and evaluate*. This is reflected in the distribution across these areas of the subset of relatively easy items. Table 7 shows that there will be challenges in covering the PISA reading domain by using relatively easy items only. In particular the test will likely be skewed a little towards access and retrieve and may contain limited reflect and evaluate material,

While this may well limit capacity to report sub-scales the implications for reading overall are less clear and in fact may not be major. Historically, in PISA, the correlations between these sub-scales have been quite high so altering the balance might be acceptable within the developing country context.

From Table 8, we see that the relatively easy pool of items for mathematics shows a modest skew towards the *Quantity* content reporting scale, but it does not appear to be problematic. For the process scales, as reported in Table 9, however, we see too few *formulate* items in the easy subset.

For science, Table 10 and Table 11 show that the reporting elements of the framework are well covered by the easy items.

Table 8. Number of secure mathematics items broken down by content sub-scales

	Number of secure items	Proportion of total	Number of relatively *easy* secure items	Proportion of total for relatively *easy* secure items	Target proportion in most recent framework
Change and relationships	28	27	8	25%	25%
Quantity	24	23	11	35%	25%
Space and shape	28	27	8	25%	25%
Uncertainty and data	25	24	5	16%	25%
Total	105		32		

Table 9. Number of secure mathematics items broken down by process sub-scales

	Number of secure items	Proportion of total	Number of relatively *easy* secure items	Proportion of total for relatively *easy* secure items	Target proportion in most recent framework
Employ	49	47%	18	56%	50%
Formulate	25	24%	3	9%	25%
Interpret	31	30%	11	34%	25%
Total	105		32		

Table 10. Number of secure science items broken down by process and content sub-scales

		Number of secure items	Proportion of total	Number of relatively *easy* secure items	Proportion of total for relatively *easy* secure items	Target proportion in most recent framework
Knowledge of science	Earth and space systems	10	11%	6	13%	12%
	Living systems	16	18%	7	16%	16%
	Physical systems	20	23%	12	27%	13%
	Technology systems	8	9%	4	9%	9%
Knowledge about science	Scientific enquiry	16	18%	8	18%	23%
	Scientific explanations	18	20%	8	18%	27%
	Total	88		45		

Table 11. Number of secure science items broken down by competency

	Number of secure items	Proportion of total	Number of relatively easy secure items	Proportion of total for relatively easy secure items	Target proportion in most recent framework
Explaining phenomena scientifically	39	44%	20	44%	41%
Identifying scientific issues	15	17%	8	18%	23%
Using scientific evidence	34	39%	17	38%	37%
Total	105		32		

In addition to the reporting scales another key aspect of the framework is item type. Item type has well-known implications for student performance and it has operational implications as well. Table 12 shows variations that reflect the relationship between item type and difficulty and suggest that it will be difficult to precisely match the framework in building a test from secure material.

Table 12. Number of secure items broken down by item format

	Reading			Mathematics		Science	
	Number (and %) of secure items	Number (and %) of *easy1* secure items	Number (and %) of *easy2* secure items	Number (and %) of secure items	Number (and %) of *easy* secure items	Number (and %) of secure items	Number (and %) of *easy* secure items
Simple multiple choice	51 (36%)	21 (57%)	31 (42%)	23 (22%)	7 (22%)	31 (35%)	23 (51%)
Auto-coded non-multiple choice	12 (8%)	0 (0%)	1 (1%)	28 (27%)	10 (31%)	25 (28%)	14 (31%)
Constructed response manual	28 (20%)	12 (32%)	21 (29%)	24 (23%)	13 (41%)	5(6%)	6 (13%)
Constructed response expert	52 (36%)	4 (11%)	20 (27%)	30 (29%)	2 (6%)	27 (31%)	2 (4%)
Total	143	37	73	105	32	88	45

> Observation 8: Drawing upon easy items only it appears that test designers will face challenges in building a test that matches the framework specifications. The implications in terms of preparing an assessment that is for purpose may not be profound, but it does suggest that it will not be possible to report at the subscale level.

For each of the three domains the issues of item difficulty range, framework coverage, item-response types, will need to be considered carefully. Working only with easier items will not guarantee adequate coverage of these areas. These issues also need to be considered in terms of each country's perception of the relevance of different parts of the framework. This question cannot be answered until initial contact and assessment of a country's needs is undertaken. This will have implications for analysis and reporting

Difficulty of selections of targeted items

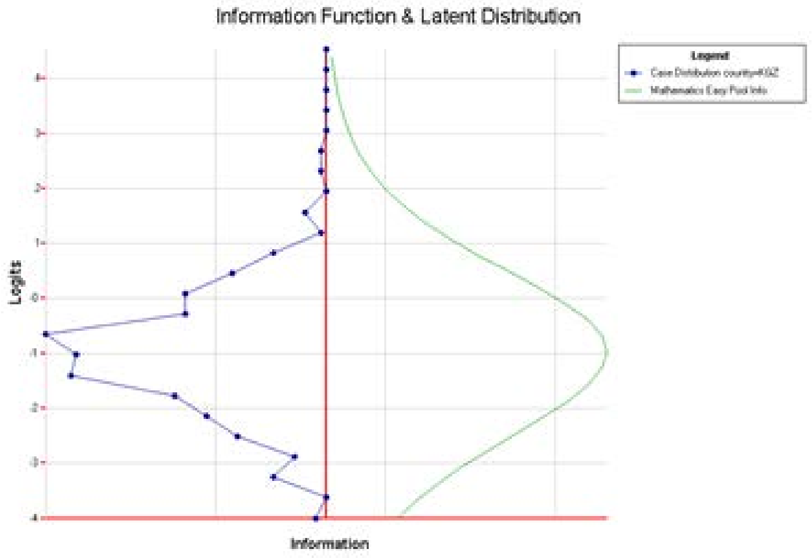
Earlier we showed that the complete item pool was too difficult for use in some contexts. Above we have considered a minimal selection of the easiest items and examined the feasibility of developing a test from them that covered the framework tolerably well. The question remains however, whether these easier subsets of items are sufficiently well targeted. As mathematics is the most challenging domain we compare the information provided by the easy mathematics items with the proficiency distribution for Kyrgyzstan.

Table 13. Proportion of total information available for each 10th of the Kyrgyzstan proficiency distribution available from the easy mathematics items

	Interval									
	Less than -2.55	-2.55 to -2.12	-2.12 to -1.91	-1.91 to -1.59	-1.59 to -1.27	-1.27 to -1.06	-1.06 to -0.74	-0.74 to -0.42	-0.42 to -0.11	Greater than -0.11
Proportion of information	0.21	0.08	0.04	0.07	0.07	0.06	0.06	0.08	0.06	0.27

The comparison, shown in Table 13 and Figure 7, illustrates a dramatic improvement when compared to Table 3and Figure 6 although we see that even the easiest pool of items that approximated the framework is more difficult than would be ideal.

Figure 7. Comparison of the item pool information function for 32 easiest mathematics items with Kyrgyzstan proficiency distribution



Observation 9: If an easy subset of items that approximates the framework is selected from the secure pool it will remain more difficult than is psychometrically ideal for many developing countries – *i.e.* with the smallest possible measurement error. In other words the test will be mis-targeted.

Countries will also be part of the normal PISA item review process where national centres are given the opportunity to view the items and comment on their perceived cultural bias and if the material in the items is part of the country's curriculum.

TEST DESIGN

PISA for Development test design

The test which will be administered to students of countries participating in the PISA for Development project should give a precise measure of the students' strengths and weaknesses so that policy makers and practitioners can implement changes which will lead to long term improvement in student capacity. At the same time, the results from PISA for Development should be directly comparable with those from PISA.

To improve the psychometric properties of the test and target it for PISA for Development it would seem appropriate to omit some of the more difficult of the secure items. Whilst we recognise that there may be a demand to include the same proportion of items from the more difficult end of the scale as in the regular PISA survey this is unnecessary and in a large number of cases result in a negative test/learning experience for the students. Through the use of appropriate modelling it is possible to reliably forecast from a targeted test the proportion of participating students who would successfully complete the most difficult items. The value and validity of an assessment is enhanced if it targets the likely performance levels of students, rather than aspirational levels.

Overview of the features of the current PISA design

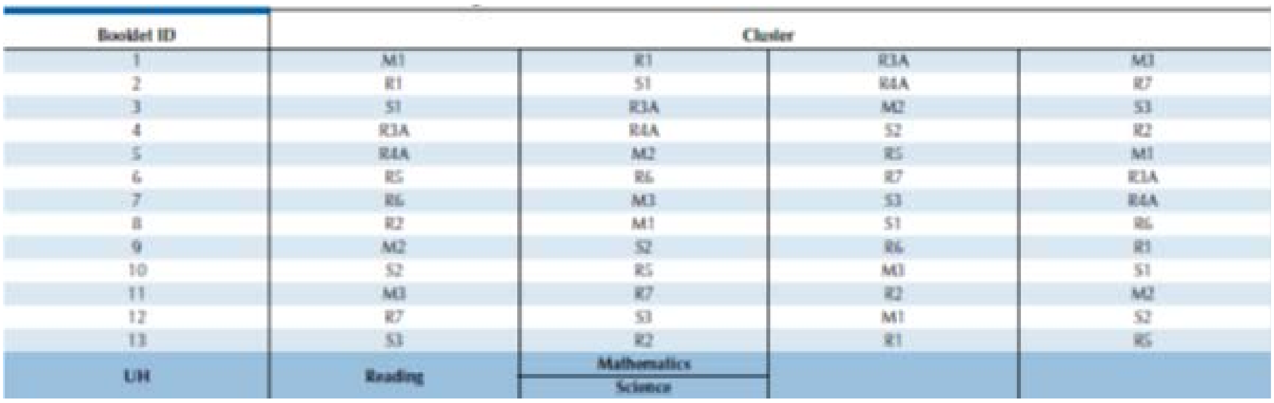
The current PISA test design is a response to many different constraints which include:

- The necessity to include items with a wide range of difficulty levels.
- The desire to have a cognitive test which takes two hours.
- The practice of having one subject area (domain) which is the focus for each three-yearly survey - starting with reading in PISA 2000, mathematics in PISA 2003 and science in PISA 2006. The next PISA survey in 2015 has science as its focus for the second time.

These constraints have resulted in a test design which has the following characteristics and associated implications:

- The assessment is made up of a two-hour cognitive test and a 30 minute questionnaire.
- The test is based on a rotated booklet design with 13 booklets. The design for PISA 2009 is shown in Figure 8. It can be seen in Figure 8 that each booklet is made up of four clusters of items from the subjects reading, mathematics and science. There are seven reading clusters (R1, R2, R3A, R4A, R5, R6, R7), three mathematics clusters (M1, M2, M3) and three science clusters (S1, S2, S3). Because reading was the major domain in PISA 2009, there is a cluster of reading in every booklet.

Figure 8. Cluster rotation design used to form standard test booklets for PISA 2009

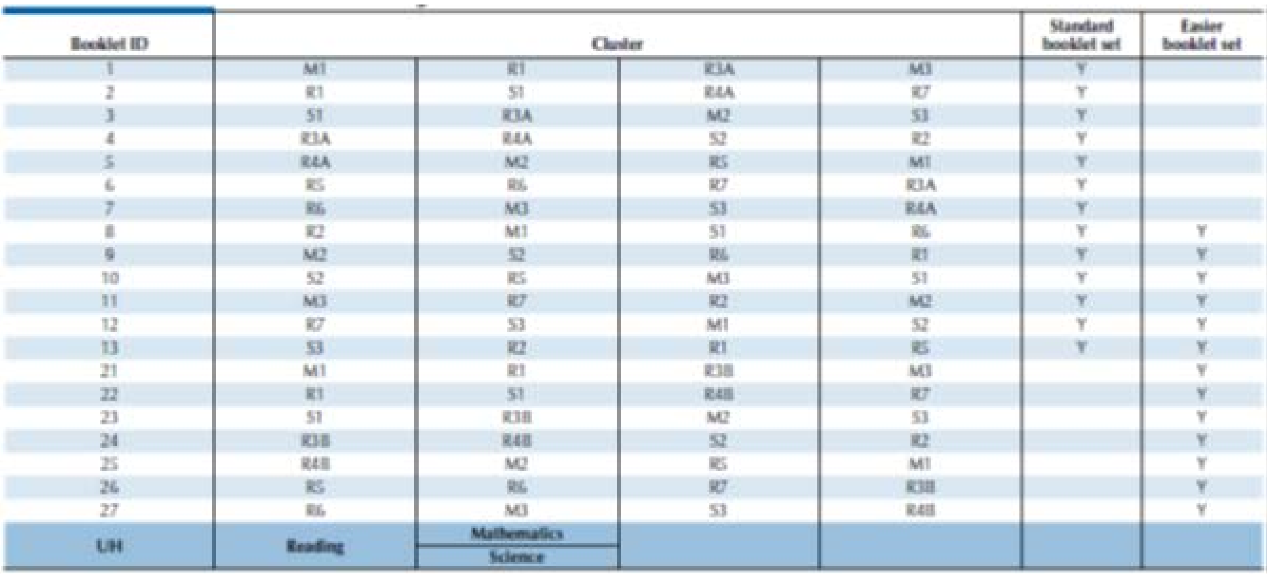


Booklet ID	Cluster			
1	M1	R1	R3A	M3
2	R1	S1	R4A	R7
3	S1	R3A	M2	S3
4	R3A	R4A	S2	R2
5	R4A	M2	R5	M1
6	R5	R6	R7	R3A
7	R6	M3	S3	R4A
8	R2	M1	S1	R6
9	M2	S2	R6	R1
10	S2	R5	M3	S1
11	M3	R7	R2	M2
12	R7	S3	M1	S2
13	S3	R2	R1	R5
UH	Reading	Mathematics Science		

Source: OECD, 2011.

In response to the need to further develop the described proficiency levels, particularly at the lower end of student capacity; four new clusters for reading were developed for PISA 2009. The rotation of the standard booklet set and the easier booklet set can be seen in Figure 9. Countries were offered the option of using the easier booklet set based on their previous PISA results.

Figure 9. Cluster rotation design used to form all test booklets for PISA 2009



Booklet ID	Cluster				Standard booklet set	Easier booklet set
1	M1	R1	R3A	M3	Y	
2	R1	S1	R4A	R7	Y	
3	S1	R3A	M2	S3	Y	
4	R3A	R4A	S2	R2	Y	
5	R4A	M2	R5	M1	Y	
6	R5	R6	R7	R3A	Y	
7	R6	M3	S3	R4A	Y	
8	R2	M1	S1	R6	Y	Y
9	M2	S2	R6	R1	Y	Y
10	S2	R5	M3	S1	Y	Y
11	M3	R7	R2	M2	Y	Y
12	R7	S3	M1	S2	Y	Y
13	S3	R2	R1	R5	Y	Y
21	M1	R1	R3B	M3		Y
22	R1	S1	R4B	R7		Y
23	S1	R3B	M2	S3		Y
24	R3B	R4B	S2	R2		Y
25	R4B	M2	R5	M1		Y
26	R5	R6	R7	R3B		Y
27	R6	M3	S3	R4B		Y
UH	Reading	Mathematics Science				

Source: OECD, 2011.

Some clusters are repeated in different booklets so that links can be made between booklets ensuring that a scale can be constructed from all available items.

Each booklet is made up of four clusters of test items – each of which takes students around 15 minutes to complete.

The clusters and consequently the booklets are designed to be of approximately equal difficulty.

Each test is administered by a trained test administrator.

Testing is carried out in a six-week window in a period not close to the start of the academic year.

In addition to the thirteen two-hour booklets, a special one-hour booklet, referred to as the UH Booklet (*Une Heure* booklet), was prepared for use in schools catering for students with special needs. The UH Booklet contained about half as many items as the other booklets, with about 50% of the items being reading items, 25% mathematics and 25% science. The items were selected from the main survey items taking into account their suitability for students with special educational needs.

Having 13 different booklets makes booklet construction a very time consuming task. Each item must be proofread to ensure exact match to the original, each item must have exactly the same layout so that the font size matches the verified translated source item, and the bolding and italicisation are consistent with the verified translated source item. Although the same item may appear in a number of booklets, it is essential that each one of them is checked in each booklet.

The need to recruit trained independent test administrators can also be a difficult task. Often there are very few people with the necessary qualifications available at the time of testing.

We understand that for PISA 2015 an even more complicated design is being proposed with some of the logistic difficulties of its implementation being offset by a centralised approach to booklet construction.

Possible test design

If the PISA for Development assessment uses a targeting process to exclude some of the more difficult items, then it is possible that there will need to be a variation to the 13 booklet rotated design used in regular PISA.

With a reduced number of items there will be no need to prepare as many item clusters.

One such design, for example is a rotation scheme consisting of three booklets. This would be very effective if one hour’s worth of suitable secure testing material per domain was identified, and building booklets of 2-hours in duration, a design like that shown in Table 14 would be suitable.

Table 14. Possible test design

Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
One	M_1	M_2	S_1	S_2
Two	S_2	S_1	R_1	R_2
Three	R_2	R_1	M_2	M_1

This design uses one hour’s worth of testing material for each domain and the booklets are two hours long, the same length as PISA. Some implications and observations from this design are as follows:

- There is no major domain, that is, all three assessment domains are equally represented.

- The two clusters for each of the domains would be constructed by removing the most difficult items from the secure material. As discussed earlier, this will provide reasonable coverage of the frameworks, but it will not permit coverage of the sub-scales.

- A shorter booklet, *i.e.* less than two hours has not been suggested because of the detrimental impact of such a change on comparability. This impact was demonstrated in analysis of earlier PISA surveys by the Technical Advisory Group.

- For the purposes of school testing we would see no difficulty in randomly selecting from one of the above three booklets or with using a separate one-hour booklet similar to the current UH booklet.

- For PISA the use of just one-hour of material (two clusters) has been deemed to be insufficient for sufficiently stable trends, especially for reading where the unit structure has the largest impact on item independence, effectively decreasing the item sample size.

- If an increased amount of testing time per domain was seen as essential then three clusters per domain would yield the same amount of testing time per domain as in the regular PISA survey - 90 minutes.

  Observation 10: A rotation scheme with fewer booklets than is used in regular PISA will be sufficient due to the limited pool of suitable testing material.

Experience with PISA has shown that even modest changes in test design can have an impact on the observed outcomes. For example, in PISA 2009 a set of easier test booklet was introduced for use in countries with a mean PISA scale score in reading of less than 440. Similarly, a number of detailed analyses of PISA data have shown than the difficulty of preceding items influences the difficultly of subsequent items. Given the likely differences between the PISA for Development design and those currently used in PISA it would therefore seem wise to implement an equating study between the PISA instruments and the PISA for development instruments so as to support the equivalence of the results.

> Observation 11: Without access to computer-delivery and working within the constraints of the available secure item pool a simplified test rotation design could be developed to provide domain level, but not sub-domain level proficiency estimates. The issue of availability of computer-delivery is central to the PISA for Development project. If tests are delivered by computer the PISA 2015 items will be used as a model, while if computer facilities are not to be used then paper-based items will form the basis of the test. This applies to all three domains.

> Observation 12: The limited size of the available secure pool will add instability to trends over time and the comparability of the results across countries. A possible way to increase the size of the item pool will be to use good easier items that have been rejected in previous PISA surveys after the field trial. They may have had good item characteristics but were rejected for other reasons - perhaps to ensure an adequate balance of item difficulty or because the framework was sufficiently covered by other items. No new development would be needed for these items.

> Observation 13: The requirement of testing 15-year-olds who are not in school has major implications for the test design. It is likely that a separate test will be designed for these students. The possibility of delivering the test by computer can also be considered using a model similar to that used in the PIAAC project. In addition the level of the test will need to be carefully considered. The PISA Reading Components assessment may be more appropriate to administer to out-of-school students. This has the disadvantage of testing reading only and lacking a real link to PISA because only two countries have participated in this assessment.

> Observation 14: It will be essential to implement an equating study between the PISA instruments and the PISA for development instruments so as to document the comparability.

Target population

For regular PISA, the desired base target population in each country consists of 15-year-old students attending educational institutions in grades 7 and higher. This means that countries include:

- 15-year-olds enrolled full-time in educational institutions;
- 15-year-olds enrolled in educational institutions who attend only on a part-time basis;
- students in vocational training programmes, or any other related type of educational programmes; and
- students attending foreign schools within the country (as well as students from other countries attending any of the programmes in the first three categories).

Whenever strict comparability with regular PISA results is needed, it will be necessary for the PISA for Development project to use the same definition. However, should a country wish to conduct the assessment with a broader range of students then this could be catered for also – for example, it may be found that there is a significant proportion of 15-year-olds who are in grades 6 or 5. At the same time, it may also be found that there is a significant proportion of students in the modal grade (*i.e.* the grade with most 15-year-olds) that are not 15 years of age – countries may wish to test these students to better understand their capabilities.

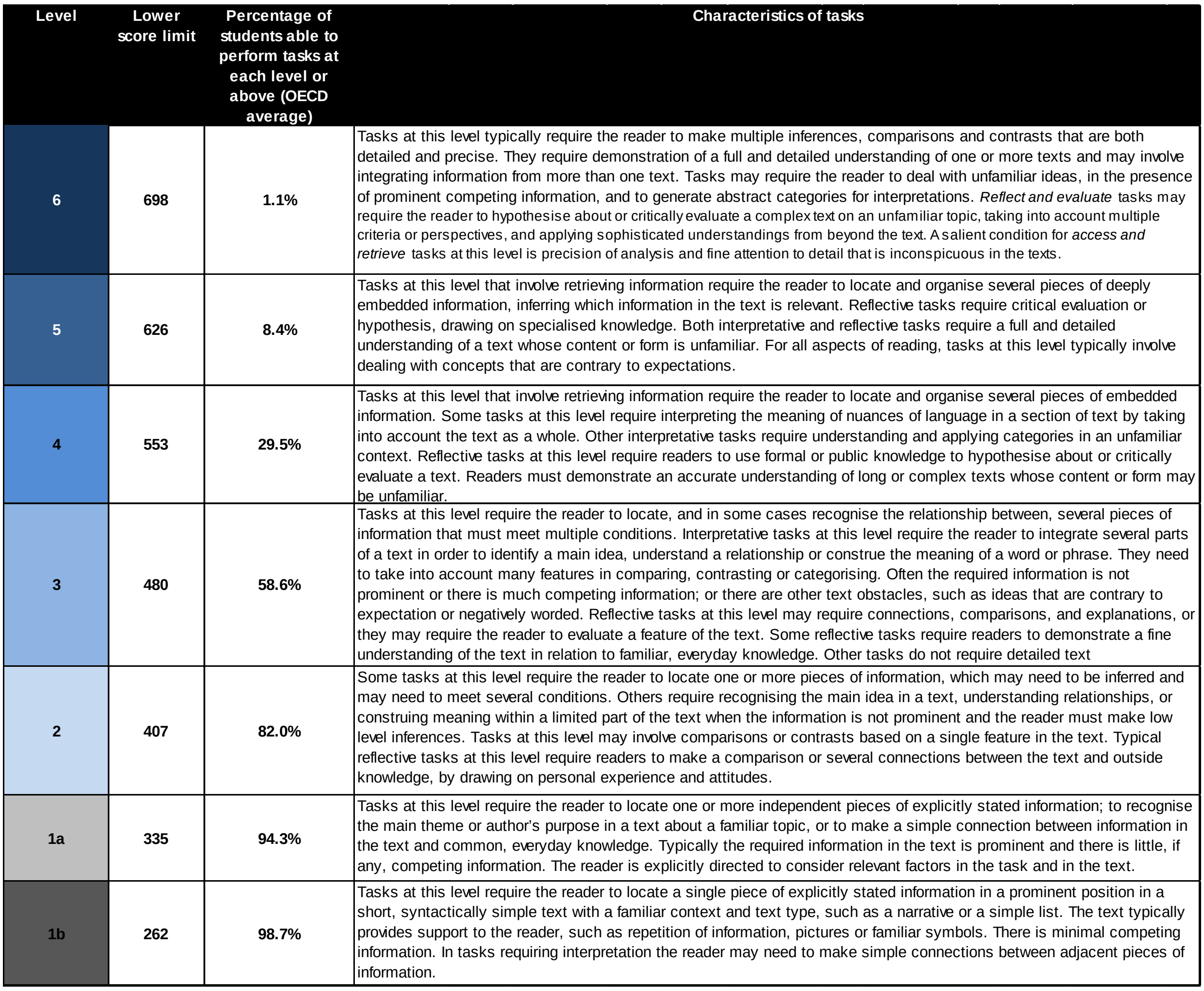
REVIEW OF THE PROFICIENCY LEVELS

PISA reading proficiency levels

The main aim of PISA is to describe the capacities of students of different education systems. While mean scores give an indication of one system's results compared to another they do little to convey what students can and cannot do. To do this PISA carefully relates the levels of difficulty of different items to students' ability to answer those items. Each item has a clearly defined purpose so it becomes easy to deduce what the students are capable of. For example in reading there are seven described levels as shown in Figure 10 (OECD, 2014). Originally, PISA reading proficiency was described in five levels, but in response to the participation of a growing number of diverse countries it was decided to increase the number of proficiency levels in reading to seven, by dividing level 1 into two parts – 1a and 1b – and to add level 6.

Table 15 shows the level number, the lower level score of the level, the percentage of students across the OECD who achieve this level and the capacity of the students at that level described in terms of the tasks that they could achieve.

Figure 10. PISA proficiency levels in reading



Level	Lower score limit	Percentage of students able to perform tasks at each level or above (OECD average)	Characteristics of tasks
6	698	1.1%	Tasks at this level typically require the reader to make multiple inferences, comparisons and contrasts that are both detailed and precise. They require demonstration of a full and detailed understanding of one or more texts and may involve integrating information from more than one text. Tasks may require the reader to deal with unfamiliar ideas, in the presence of prominent competing information, and to generate abstract categories for interpretations. *Reflect and evaluate* tasks may require the reader to hypothesise about or critically evaluate a complex text on an unfamiliar topic, taking into account multiple criteria or perspectives, and applying sophisticated understandings from beyond the text. A salient condition for *access and retrieve* tasks at this level is precision of analysis and fine attention to detail that is inconspicuous in the texts.
5	626	8.4%	Tasks at this level that involve retrieving information require the reader to locate and organise several pieces of deeply embedded information, inferring which information in the text is relevant. Reflective tasks require critical evaluation or hypothesis, drawing on specialised knowledge. Both interpretative and reflective tasks require a full and detailed understanding of a text whose content or form is unfamiliar. For all aspects of reading, tasks at this level typically involve dealing with concepts that are contrary to expectations.
4	553	29.5%	Tasks at this level that involve retrieving information require the reader to locate and organise several pieces of embedded information. Some tasks at this level require interpreting the meaning of nuances of language in a section of text by taking into account the text as a whole. Other interpretative tasks require understanding and applying categories in an unfamiliar context. Reflective tasks at this level require readers to use formal or public knowledge to hypothesise about or critically evaluate a text. Readers must demonstrate an accurate understanding of long or complex texts whose content or form may be unfamiliar.
3	480	58.6%	Tasks at this level require the reader to locate, and in some cases recognise the relationship between, several pieces of information that must meet multiple conditions. Interpretative tasks at this level require the reader to integrate several parts of a text in order to identify a main idea, understand a relationship or construe the meaning of a word or phrase. They need to take into account many features in comparing, contrasting or categorising. Often the required information is not prominent or there is much competing information; or there are other text obstacles, such as ideas that are contrary to expectation or negatively worded. Reflective tasks at this level may require connections, comparisons, and explanations, or they may require the reader to evaluate a feature of the text. Some reflective tasks require readers to demonstrate a fine understanding of the text in relation to familiar, everyday knowledge. Other tasks do not require detailed text
2	407	82.0%	Some tasks at this level require the reader to locate one or more pieces of information, which may need to be inferred and may need to meet several conditions. Others require recognising the main idea in a text, understanding relationships, or construing meaning within a limited part of the text when the information is not prominent and the reader must make low level inferences. Tasks at this level may involve comparisons or contrasts based on a single feature in the text. Typical reflective tasks at this level require readers to make a comparison or several connections between the text and outside knowledge, by drawing on personal experience and attitudes.
1a	335	94.3%	Tasks at this level require the reader to locate one or more independent pieces of explicitly stated information; to recognise the main theme or author's purpose in a text about a familiar topic, or to make a simple connection between information in the text and common, everyday knowledge. Typically the required information in the text is prominent and there is little, if any, competing information. The reader is explicitly directed to consider relevant factors in the task and in the text.
1b	262	98.7%	Tasks at this level require the reader to locate a single piece of explicitly stated information in a prominent position in a short, syntactically simple text with a familiar context and text type, such as a narrative or a simple list. The text typically provides support to the reader, such as repetition of information, pictures or familiar symbols. There is minimal competing information. In tasks requiring interpretation the reader may need to make simple connections between adjacent pieces of information.

Source: OECD, 2014.

Table 15. PISA 2012 percentage of students in proficiency levels in reading by country

	% Below 1b	% 1b	% 1a	% 2	% 3	% 4	% 5	% 6
Australia	0.9	3.1	10.2	21.6	29.1	23.3	9.8	1.9
Austria	0.8	4.8	13.8	24.2	29.6	21.2	5.2	0.3
Belgium	1.6	4.1	10.4	20.4	27.3	24.4	10.4	1.4
Canada	0.5	2.4	8.0	19.4	31.0	25.8	10.8	2.1
Chile	1.0	8.1	23.9	35.1	24.3	6.9	0.6	0.0
Czech Republic	0.6	3.5	12.7	26.4	31.3	19.4	5.3	0.8
Denmark	0.8	3.1	10.7	25.8	33.6	20.5	5.1	0.4
Estonia	0.2	1.3	7.7	22.7	35.0	24.9	7.5	0.9
Finland	0.7	2.4	8.2	19.1	29.3	26.8	11.3	2.2
France	2.1	4.9	11.9	18.9	26.3	23.0	10.6	2.3
Germany	0.5	3.3	10.7	22.1	29.9	24.6	8.3	0.7
Greece	2.6	5.9	14.2	25.1	30.0	17.2	4.6	0.5
Hungary	0.7	5.2	13.8	24.3	29.9	20.4	5.3	0.4
Iceland	2.3	5.4	13.3	24.7	29.9	18.6	5.2	0.6
Ireland	0.3	1.9	7.5	19.6	33.4	26.0	10.1	1.3
Israel	3.8	6.9	12.9	20.8	25.3	20.6	8.1	1.5
Italy	1.6	5.2	12.7	23.7	29.7	20.5	6.1	0.6
Japan	0.6	2.4	6.7	16.6	26.7	28.4	14.6	3.9
Korea	0.4	1.7	5.5	16.4	30.8	31.0	12.6	1.6
Luxembourg	2.0	6.3	13.8	23.4	25.8	19.7	7.5	1.4
Mexico	2.6	11.0	27.5	34.5	19.6	4.5	0.4	0.0
Netherlands	0.9	2.8	10.3	21.0	29.2	26.1	9.0	0.8
New Zealand	1.3	4.0	11.0	20.8	26.3	22.7	10.9	3.0
Norway	1.7	3.7	10.8	21.9	29.4	22.3	8.5	1.7
Poland	0.3	2.1	8.1	21.4	32.0	26.0	8.6	1.4
Portugal	1.3	5.1	12.3	25.5	30.2	19.7	5.3	0.5
Slovak Republic	4.1	7.9	16.2	25.0	26.8	15.7	4.1	0.3
Slovenia	1.2	4.9	15.0	27.2	28.4	18.2	4.7	0.3
Spain	1.3	4.4	12.6	25.8	31.2	19.2	5.0	0.5
Sweden	2.9	6.0	13.9	23.5	27.3	18.6	6.7	1.2
Switzerland	0.5	2.9	10.3	21.9	31.5	23.8	8.2	1.0
Turkey	0.6	4.5	16.6	30.8	28.7	14.5	4.1	0.3
United Kingdom	1.5	4.0	11.2	23.5	29.9	21.3	7.5	1.3
United States	0.8	3.6	12.3	24.9	30.5	20.1	6.9	1.0
OECD average	1.3	4.4	12.3	23.5	29.1	21.0	7.3	1.1
Albania	12.0	15.9	24.4	24.7	15.9	5.9	1.1	0.1
Argentina	8.1	17.7	27.7	27.3	14.6	4.0	0.5	0.1
Brazil	4.0	14.8	30.4	30.1	15.8	(4.4)	0.5	0.0
Bulgaria	8.0	12.8	18.6	22.2	21.4	12.7	3.8	0.5
Colombia	5.0	15.4	31.0	30.5	14.5	3.2	0.3	0.0
Costa Rica	0.8	7.3	24.3	38.1	22.9	6.0	0.6	0.0
Croatia	0.7	4.0	13.9	27.8	31.2	17.8	4.2	0.2
Cyprus[2]	6.1	9.7	17.0	25.1	24.9	13.2	3.5	0.5
Hong Kong-China	0.2	1.3	5.3	14.3	29.2	32.9	14.9	1.9
Indonesia	4.1	16.3	34.8	31.6	11.5	1.5	0.1	0.0
Jordan	7.5	14.9	28.3	30.8	15.5	2.9	0.1	0.0
Kazakhstan	4.2	17.3	35.6	31.3	10.4	1.2	0.0	0.0
Latvia	0.7	3.7	12.6	26.7	33.1	19.1	3.9	0.3
Liechtenstein	0.0	1.9	10.5	22.4	28.6	25.7	10.4	0.6
Lithuania	1.0	4.6	15.6	28.1	31.1	16.3	3.1	0.2
Macao-China	0.3	2.1	9.0	23.3	34.3	24.0	6.4	0.6
Malaysia	5.8	16.4	30.5	31.0	13.6	2.5	0.1	0.0
Montenegro	4.4	13.2	25.7	29.2	19.9	6.6	0.9	0.0

[2] **Notes regarding Cyprus**

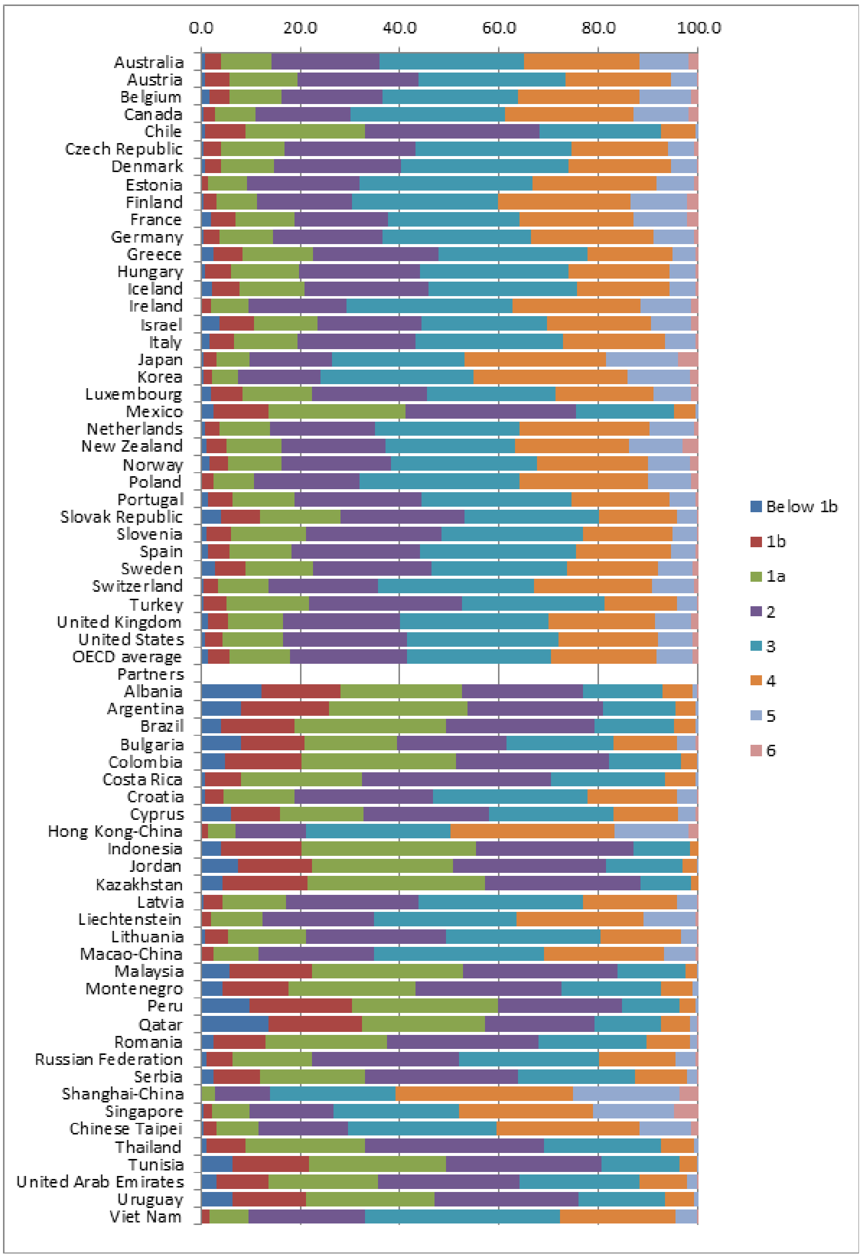
Note by Turkey: The information in this document with reference to “Cyprus” relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the “Cyprus issue”.

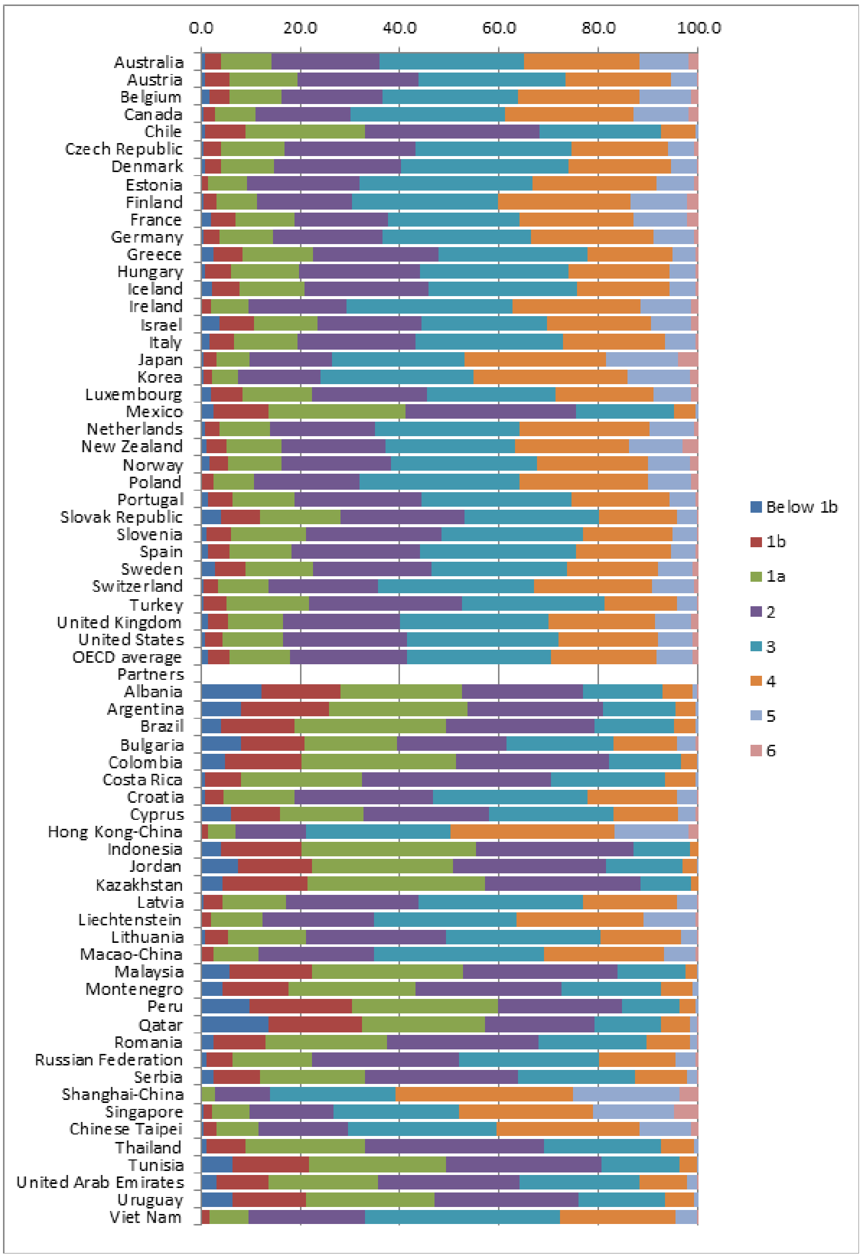
Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

	% Below 1b	% 1b	% 1a	% 2	% 3	% 4	% 5	% 6
Peru	9.8	20.6	29.5	24.9	11.4	3.3	0.5	0.0
Qatar	13.6	18.9	24.6	21.9	13.5	5.8	1.4	0.2
Romania	2.5	10.3	24.4	30.6	21.8	8.7	1.5	0.1
Russian Federation	1.1	5.2	16.0	29.5	28.3	15.3	4.2	0.5
Serbia	2.6	9.3	21.3	30.8	23.3	10.5	2.0	0.2
Shanghai-China	0.1	0.3	2.5	11.0	25.3	35.7	21.3	3.8
Singapore	0.5	1.9	7.5	16.7	25.4	26.8	16.2	5.0
Chinese Taipei	0.6	2.5	8.4	18.1	29.9	28.7	10.4	1.4
Thailand	1.2	7.7	24.1	36.0	23.5	6.7	0.8	0.1
Tunisia	6.2	15.5	27.6	31.4	15.6	3.5	0.2	0.0
United Arab Emirates	3.3	10.4	21.8	28.6	24.0	9.7	2.1	0.2
Uruguay	6.4	14.7	25.9	28.9	17.4	5.7	0.9	0.0
Viet Nam	0.1	1.5	7.8	23.7	39.0	23.4	4.2	0.4

Source: OECD, 2014.

Figure 11. PISA 2012 reading proficiency levels by country





There are four main points that can be observed from Figure 11:

- The figure shows that within the OECD there are three countries which deviate from the pattern of average percentages for each proficiency level. These countries are Chile, Mexico and Turkey which have the smallest percentages of students in levels 5 and 6 but have substantial percentages of students in the lower proficiency levels. In Mexico, over 40% of the students are below level 2, and in Chile, over 30% of the students are below level 2. Level 2 is significant because it is seen as the lowest level needed for students to make adequate progress in their future learning. This has an impact not only in reading, but in almost every other activity, because reading is at the foundation of new learning.

- In the partner countries/regions there is a much greater diversity in the pattern of percentages of students in the various proficiency levels. This diversity reflects large differences in economic status, resources committed to education, the nature of the educational programmes and cultural differences.

- Among the partner countries/region, there are a number of countries having over 50% of their students not able to reach level 2. These include Albania, Argentina, Colombia, Indonesia, Jordan, Kazakhstan, Malaysia, Peru and Qatar.

- Among the partner countries/region, there are a number of countries with a notable percentage of students who are below the lowest described level - level 1b. In can be seen in Figure 2 that this level is described as:

Tasks at this level require the reader to locate a single piece of explicitly stated information in a prominent position in a short, syntactically simple text with a familiar context and text type, such as a narrative or a simple list. The text typically provides support to the reader, such as repetition of information, pictures or familiar symbols. There is minimal competing information. In tasks requiring interpretation the reader may need to make simple connections between adjacent pieces of information.

Albania (12%) and Qatar (13.6%) are the countries with the highest percentage of students who are below level 1b, but there are nine countries where more than 5% of the students do not reach level 1b - Argentina, Bulgaria, Colombia, Cyprus[3], Jordan, Malaysia, Peru, Tunisia and Uruguay. For these 11 countries PISA is unable to describe the skills that these students have - it is not that the students do not possess skills, it is just that the PISA survey is not designed to assess and describe them. Without this information it is extremely difficult for education systems to plan remediation programmes to help these students.

> Observation 15: The current PISA described proficiency levels in reading do not provide enough useful information for many developing countries making it difficult for policy makers to identify and implement remedial interventions focused on their students' weaknesses.

To expand the description of student capacity at the lower end of the scale, the PISA Reading Components assessment could be used as a basis. This was offered to countries as part of the PISA 2012 assessment, but only two countries opted for it. The framework of reading components is made up of three main areas: Word meaning (print vocabulary), Sentence processing and Basic passage comprehension.

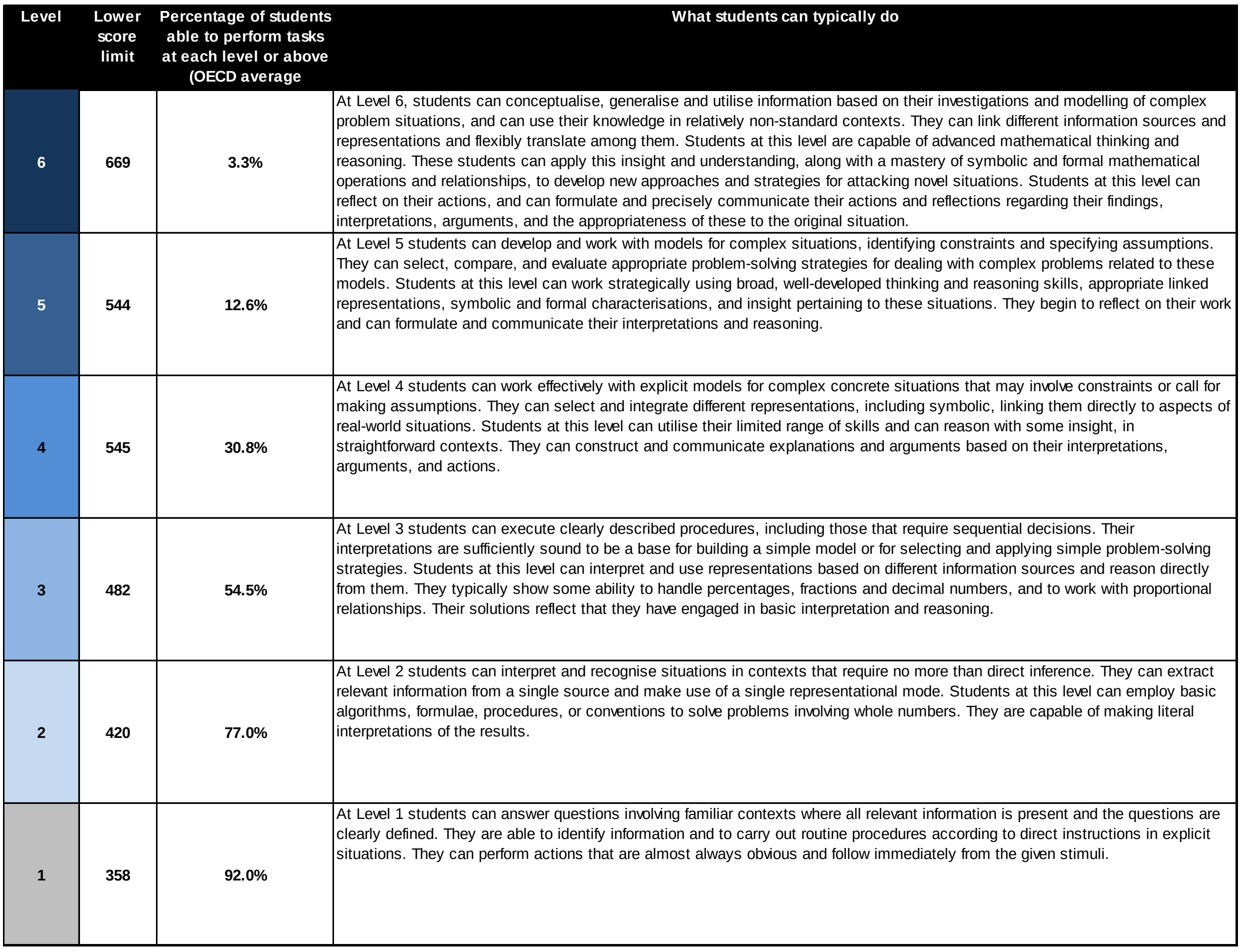
Components assessment tasks are designed to inform understanding of the basic reading skills that underlie proficient literacy performance levels. They help us describe what low ability readers can do and therefore form a basis for learning, instruction, and policy with respect to helping them achieve higher literacy levels. The PISA Reading Components assessment focused attention on those component skills that show the greatest promise for cross-country comparability, specifically reading vocabulary, sentence comprehension, and basic passage comprehension and fluency.

PISA mathematics proficiency levels

In mathematics there are six described levels as shown in Figure 12 (OECD, 2014). Table 16 shows the level number, the lower level score of the level, the percentage of students across the OECD who achieve this level and the capacity of the students at that level described in terms of the tasks that they could achieve.

[3] See footnote 2.

Figure 12. PISA proficiency levels in mathematics



Level	Lower score limit	Percentage of students able to perform tasks at each level or above (OECD average	What students can typically do
6	669	3.3%	At Level 6, students can conceptualise, generalise and utilise information based on their investigations and modelling of complex problem situations, and can use their knowledge in relatively non-standard contexts. They can link different information sources and representations and flexibly translate among them. Students at this level are capable of advanced mathematical thinking and reasoning. These students can apply this insight and understanding, along with a mastery of symbolic and formal mathematical operations and relationships, to develop new approaches and strategies for attacking novel situations. Students at this level can reflect on their actions, and can formulate and precisely communicate their actions and reflections regarding their findings, interpretations, arguments, and the appropriateness of these to the original situation.
5	544	12.6%	At Level 5 students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare, and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well-developed thinking and reasoning skills, appropriate linked representations, symbolic and formal characterisations, and insight pertaining to these situations. They begin to reflect on their work and can formulate and communicate their interpretations and reasoning.
4	545	30.8%	At Level 4 students can work effectively with explicit models for complex concrete situations that may involve constraints or call for making assumptions. They can select and integrate different representations, including symbolic, linking them directly to aspects of real-world situations. Students at this level can utilise their limited range of skills and can reason with some insight, in straightforward contexts. They can construct and communicate explanations and arguments based on their interpretations, arguments, and actions.
3	482	54.5%	At Level 3 students can execute clearly described procedures, including those that require sequential decisions. Their interpretations are sufficiently sound to be a base for building a simple model or for selecting and applying simple problem-solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They typically show some ability to handle percentages, fractions and decimal numbers, and to work with proportional relationships. Their solutions reflect that they have engaged in basic interpretation and reasoning.
2	420	77.0%	At Level 2 students can interpret and recognise situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions to solve problems involving whole numbers. They are capable of making literal interpretations of the results.
1	358	92.0%	At Level 1 students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are almost always obvious and follow immediately from the given stimuli.

Source: OECD, 2014.

Table 16. PISA 2012 percentage of students in proficiency levels in mathematics by country

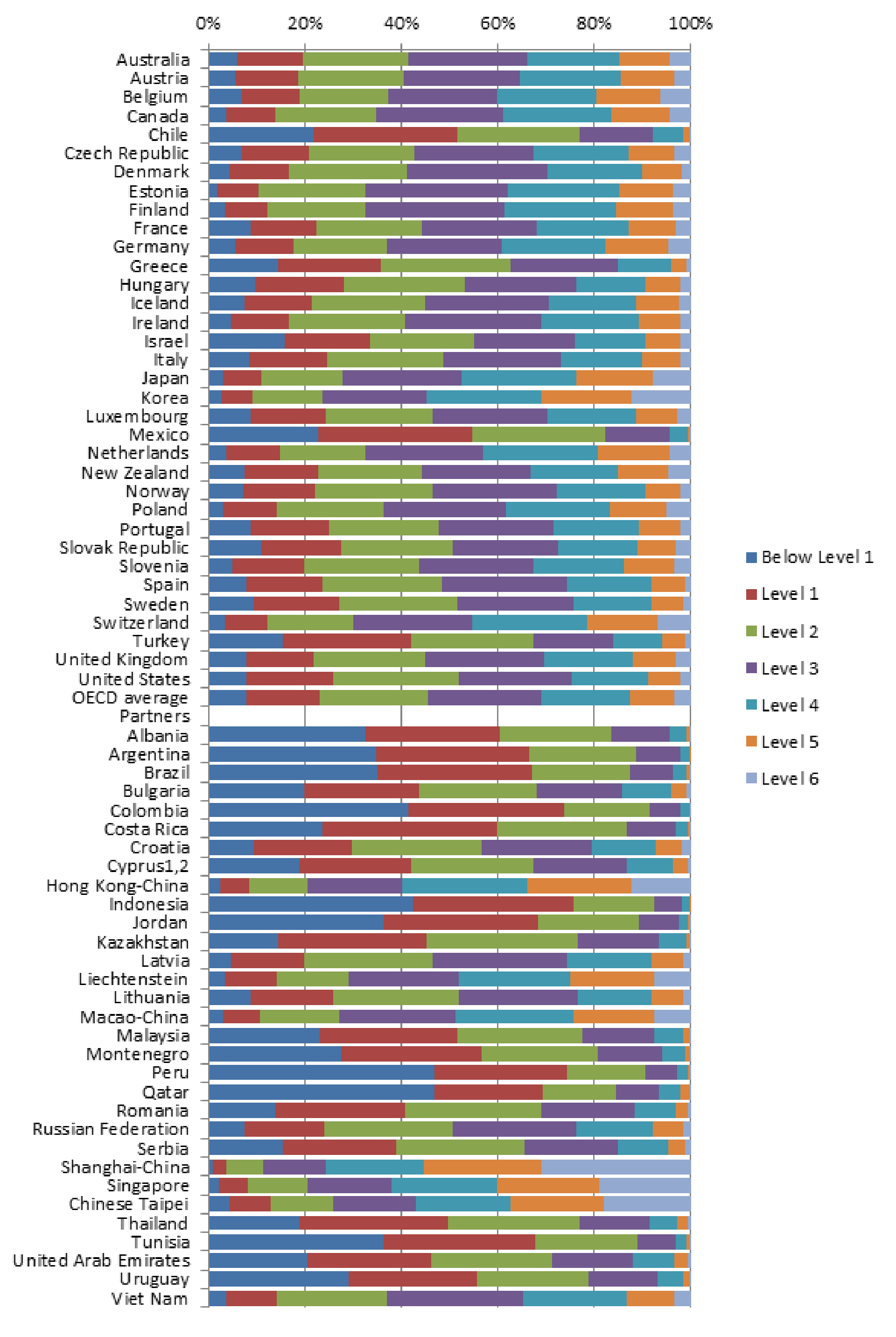
	%	%	%	%	%	%	%
	Below Level 1	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Australia	6.1	13.5	21.9	24.6	19.0	10.5	4.3
Austria	5.7	13.0	21.9	24.2	21.0	11.0	3.3
Belgium	7.0	12.0	18.4	22.4	20.6	13.4	6.1
Canada	3.6	10.2	21.0	26.4	22.4	12.1	4.3
Chile	22.0	29.5	25.3	15.4	6.2	1.5	0.1
Czech Republic	6.8	14.2	21.7	24.8	19.7	9.6	3.2
Denmark	4.4	12.5	24.4	29.0	19.8	8.3	1.7
Estonia	2.0	8.6	22.0	29.4	23.4	11.0	3.6
Finland	3.3	8.9	20.5	28.8	23.2	11.7	3.5
France	8.7	13.6	22.1	23.8	18.9	9.8	3.1
Germany	5.5	12.2	19.4	23.7	21.7	12.8	4.7
Greece	14.5	21.2	27.2	22.1	11.2	3.3	0.6
Hungary	9.9	18.2	25.3	23.0	14.4	7.1	2.1
Iceland	7.5	14.0	23.6	25.7	18.1	8.9	2.3
Ireland	4.8	12.1	23.9	28.2	20.3	8.5	2.2
Israel	15.9	17.6	21.6	21.0	14.6	7.2	2.2
Italy	8.5	16.1	24.1	24.6	16.7	7.8	2.2
Japan	3.2	7.9	16.9	24.7	23.7	16.0	7.6
Korea	2.7	6.4	14.7	21.4	23.9	18.8	12.1
Luxembourg	8.8	15.5	22.3	23.6	18.5	8.6	2.6
Mexico	22.8	31.9	27.8	13.1	3.7	0.6	0.0
Netherlands	3.8	11.0	17.9	24.2	23.8	14.9	4.4
New Zealand	7.5	15.1	21.6	22.7	18.1	10.5	4.5
Norway	7.2	15.1	24.3	25.7	18.3	7.3	2.1
Poland	3.3	11.1	22.1	25.5	21.3	11.7	5.0
Portugal	8.9	16.0	22.8	24.0	17.7	8.5	2.1
Slovak Republic	11.1	16.4	23.1	22.1	16.4	7.8	3.1
Slovenia	5.1	15.0	23.6	23.9	18.7	10.3	3.4
Spain	7.8	15.8	24.9	26.0	17.6	6.7	1.3
Sweden	9.5	17.5	24.7	23.9	16.3	6.5	1.6
Switzerland	3.6	8.9	17.8	24.5	23.9	14.6	6.8
Turkey	15.5	26.5	25.5	16.5	10.1	4.7	1.2
United Kingdom	7.8	14.0	23.2	24.8	18.4	9.0	2.9
United States	8.0	17.9	26.3	23.3	15.8	6.6	2.2
OECD total	9.1	16.9	23.3	22.2	16.5	8.6	3.3
OECD average	8.0	15.0	22.5	23.7	18.1	9.3	3.3
Albania	32.5	28.1	22.9	12.0	3.6	0.8	0.0
Argentina	34.9	31.6	22.2	9.2	1.8	0.3	0.0
Brazil	35.2	31.9	20.4	8.9	2.9	0.7	0.0
Bulgaria	20.0	23.8	24.4	17.9	9.9	3.4	0.7
Colombia	41.6	32.2	17.8	6.4	1.6	0.3	0.0
Costa Rica	23.6	36.2	26.8	10.1	2.6	0.5	0.1
Croatia	9.5	20.4	26.7	22.9	13.5	5.4	1.6
Cyprus[4]	19.0	23.0	25.5	19.2	9.6	3.1	0.6
Hong Kong-China	2.6	5.9	12.0	19.7	26.1	21.4	12.3
Indonesia	42.3	33.4	16.8	5.7	1.5	0.3	0.0
Jordan	36.5	32.1	21.0	8.1	1.8	0.5	0.1
Kazakhstan	14.5	30.7	31.5	16.9	5.4	0.9	0.1
Latvia	4.8	15.1	26.6	27.8	17.6	6.5	1.5
Liechtenstein	3.5	10.6	15.2	22.7	23.2	17.4	7.4
Lithuania	8.7	17.3	25.9	24.6	15.4	6.6	1.4
Macao-China	3.2	7.6	16.4	24.0	24.4	16.8	7.6
Malaysia	23.0	28.8	26.0	14.9	6.0	1.2	0.1
Montenegro	27.5	29.1	24.2	13.1	4.9	0.9	0.1
Peru	47.0	27.6	16.1	6.7	2.1	0.5	0.0
Qatar	47.0	22.6	15.2	8.8	4.5	1.7	0.3
Romania	14.0	26.8	28.3	19.2	8.4	2.6	0.6
Russian Federation	7.5	16.5	26.6	26.0	15.7	6.3	1.5
Serbia	15.5	23.4	26.5	19.5	10.5	3.5	1.1
Shanghai-China	0.8	2.9	7.5	13.1	20.2	24.6	30.8
Singapore	2.2	6.1	12.2	17.5	22.0	21.0	19.0
Chinese Taipei	4.5	8.3	13.1	17.1	19.7	19.2	18.0
Thailand	19.1	30.6	27.3	14.5	5.8	2.0	0.5

[4] See footnote 2.

	% Below Level 1	% Level 1	% Level 2	% Level 3	% Level 4	% Level 5	% Level 6
Tunisia	36.5	31.3	21.1	8.0	2.3	0.7	0.1
United Arab Emirates	20.5	25.8	24.9	16.9	8.5	2.9	0.5
Uruguay	29.2	26.5	23.0	14.4	5.4	1.3	0.1
Viet Nam	3.6	10.6	22.8	28.4	21.3	9.8	3.5

Source: OECD, 2014.

Figure 13. PISA 2012 mathematics proficiency levels by country



The main points that can be observed from Figure 12 are:

- The figure shows that within the OECD there are two countries which deviate from the pattern of average percentages for each proficiency level. These countries are Chile and Mexico which have the smallest percentages of students in levels 5 and 6, but have substantial percentages of students in the lower proficiency levels. In both countries there are over 50% of the students who are below level 2.

- In a manner similar to the results in reading, in the partner countries/regions there is a much greater diversity in the pattern of percentages of students in the various proficiency levels. This diversity reflects large differences in economic status, resources committed to education, the nature of the educational programmes and cultural differences.

- Among the partner countries/region, there is a number of countries having over 60% of their students not able to reach level 2. These include Albania, Argentina, Brazil, Colombia, Indonesia, Jordan, Peru, Qatar and Tunisia.

- Among the partner countries/region, there is a number of countries with a notable percentage of students who are below the lowest described level – level 1. In can be seen in Figure 3 that this level is described as:

At Level 1, students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are almost always obvious and follow immediately from the given stimuli.

- Colombia (42%), Indonesia (42%), Peru and Qatar (both 47%) are the countries with the highest percentage of students who are below level 1, but there are five countries where more than 30% of the students do not reach level 1 – Albania, Argentina, Brazil, Jordan, and Tunisia. For these nine countries PISA is unable to describe the skills that these students have - it is not that the students do not possess skills, it is just that the PISA survey is not designed to assess and describe them. Without this information it is extremely difficult for education systems to plan remediation programmes to help these students.

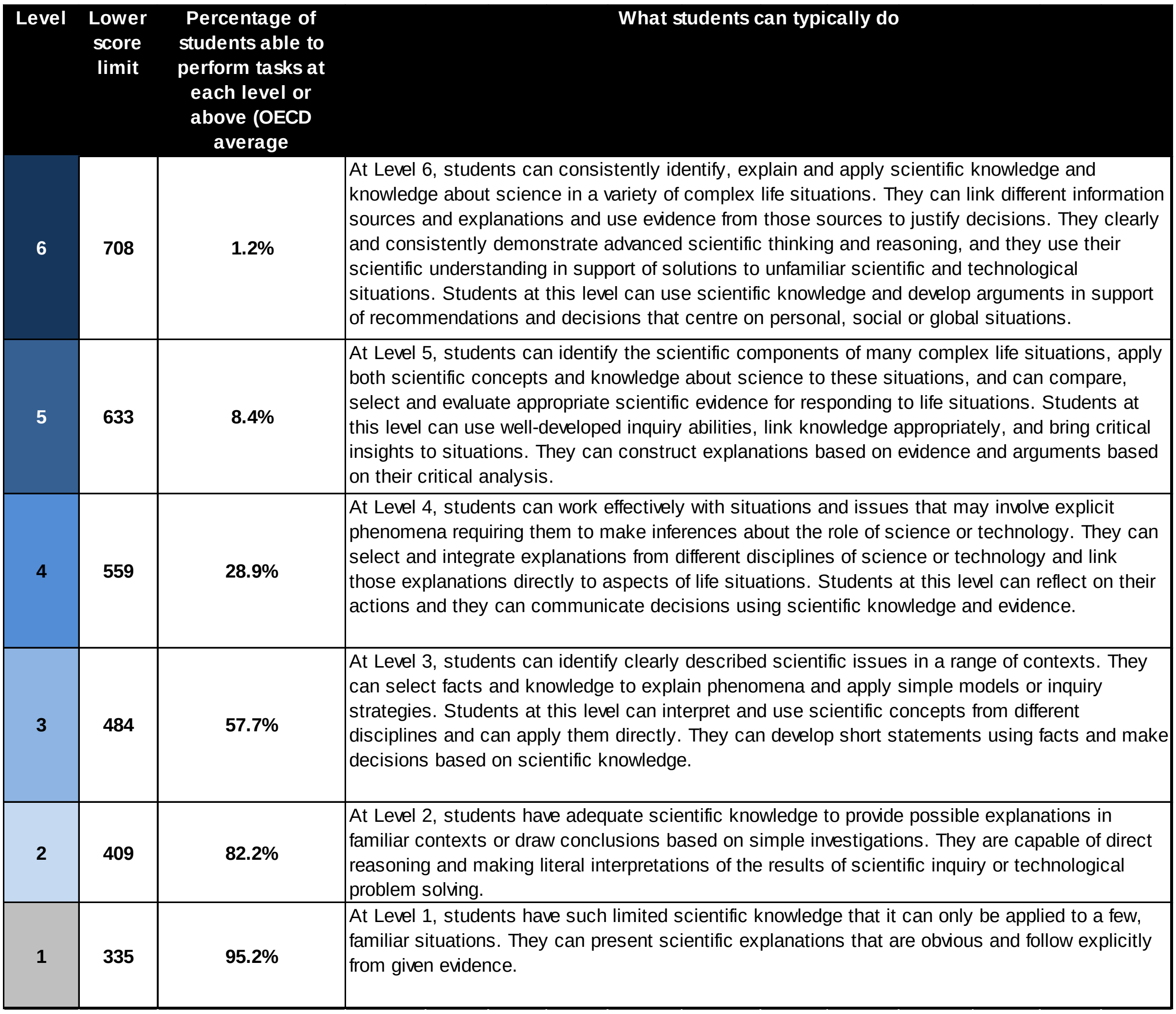
  Observation 16: In mathematics, in some countries, nearly half the students are below the lowest level for which PISA can describe student capacity.

PISA has not undertaken a components assessment of mathematics as was done with Reading Components. While it could be possible to explore assessments such as Early Grade Mathematics Assessment (EGMA) the amount of development required would preclude this step from the PISA for Development pilot project. EGMA is an oral assessment designed to measure a student's foundation skills in numeracy and mathematics in the early grades, including number identification, quantity discrimination, missing-number identification, word problem solving, addition and subtraction, shape recognition, and pattern extension. The EGMA instrument is adapted for use in a particular country and language.

PISA science proficiency levels

In science there are six described levels as shown in Figure 14 (OECD, 2014). Figure 5 shows the level number, the lower level score of the level, the percentage of students across the OECD who achieve this level and the capacity of the students at that level described in terms of the tasks that they could achieve.

Figure 14. PISA proficiency levels in science



Level	Lower score limit	Percentage of students able to perform tasks at each level or above (OECD average	What students can typically do
6	708	1.2%	At Level 6, students can consistently identify, explain and apply scientific knowledge and knowledge about science in a variety of complex life situations. They can link different information sources and explanations and use evidence from those sources to justify decisions. They clearly and consistently demonstrate advanced scientific thinking and reasoning, and they use their scientific understanding in support of solutions to unfamiliar scientific and technological situations. Students at this level can use scientific knowledge and develop arguments in support of recommendations and decisions that centre on personal, social or global situations.
5	633	8.4%	At Level 5, students can identify the scientific components of many complex life situations, apply both scientific concepts and knowledge about science to these situations, and can compare, select and evaluate appropriate scientific evidence for responding to life situations. Students at this level can use well-developed inquiry abilities, link knowledge appropriately, and bring critical insights to situations. They can construct explanations based on evidence and arguments based on their critical analysis.
4	559	28.9%	At Level 4, students can work effectively with situations and issues that may involve explicit phenomena requiring them to make inferences about the role of science or technology. They can select and integrate explanations from different disciplines of science or technology and link those explanations directly to aspects of life situations. Students at this level can reflect on their actions and they can communicate decisions using scientific knowledge and evidence.
3	484	57.7%	At Level 3, students can identify clearly described scientific issues in a range of contexts. They can select facts and knowledge to explain phenomena and apply simple models or inquiry strategies. Students at this level can interpret and use scientific concepts from different disciplines and can apply them directly. They can develop short statements using facts and make decisions based on scientific knowledge.
2	409	82.2%	At Level 2, students have adequate scientific knowledge to provide possible explanations in familiar contexts or draw conclusions based on simple investigations. They are capable of direct reasoning and making literal interpretations of the results of scientific inquiry or technological problem solving.
1	335	95.2%	At Level 1, students have such limited scientific knowledge that it can only be applied to a few, familiar situations. They can present scientific explanations that are obvious and follow explicitly from given evidence.

Source: OECD, 2014.

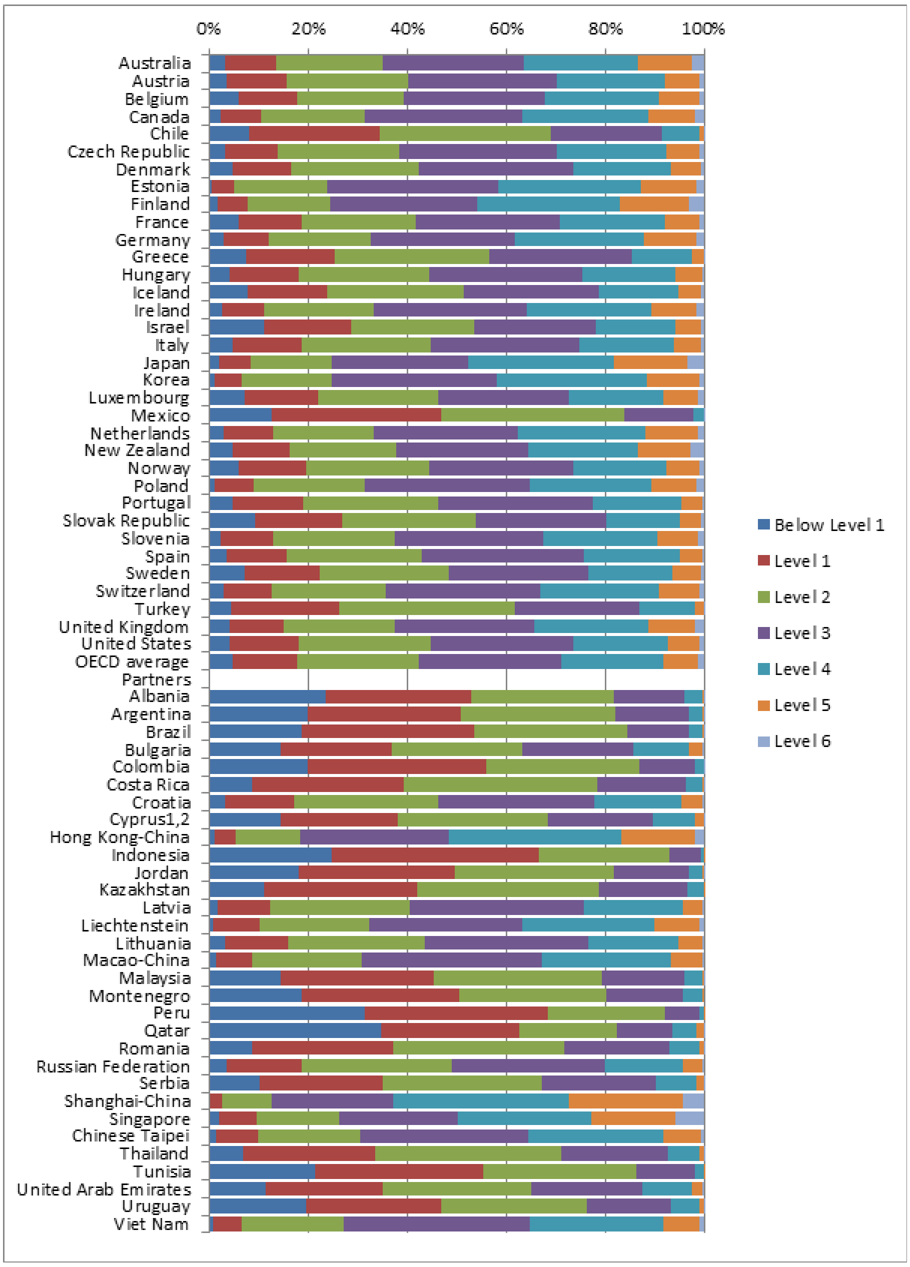
Table 17. PISA 2012 percentage of students in proficiency in science by country

	% Below Level 1	% Level 1	% Level 2	% Level 3	% Level 4	% Level 5	% Level 6
Australia	3.4	10.2	21.5	28.5	22.8	10.9	2.6
Austria	3.6	12.2	24.3	30.1	21.9	7.0	0.8
Belgium	5.9	11.8	21.5	28.7	23.0	8.1	0.9
Canada	2.4	8.0	21.0	32.0	25.3	9.5	1.8
Chile	8.1	26.3	34.6	22.4	7.5	1.0	0.0
Czech Republic	3.3	10.5	24.7	31.7	22.2	6.7	0.9
Denmark	4.7	12.0	25.7	31.3	19.6	6.1	0.7
Estonia	0.5	4.5	19.0	34.5	28.7	11.1	1.7
Finland	1.8	5.9	16.8	29.6	28.8	13.9	3.2
France	6.1	12.6	22.9	29.2	21.3	6.9	1.0
Germany	2.9	9.3	20.5	28.9	26.2	10.6	1.6
Greece	7.4	18.1	31.0	28.8	12.2	2.3	0.2
Hungary	4.1	14.0	26.4	30.9	18.7	5.5	0.5
Iceland	8.0	16.0	27.5	27.2	16.2	4.6	0.6
Ireland	2.6	8.5	22.0	31.1	25.0	9.3	1.5
Israel	11.2	17.7	24.8	24.4	16.1	5.2	0.6
Italy	4.9	13.8	26.0	30.1	19.1	5.5	0.6
Japan	2.0	6.4	16.3	27.5	29.5	14.8	3.4
Korea	1.2	5.5	18.0	33.6	30.1	10.6	1.1
Luxembourg	7.2	15.1	24.2	26.2	19.2	7.0	1.2
Mexico	12.6	34.4	37.0	13.8	2.1	0.1	0.0
Netherlands	3.1	10.1	20.1	29.1	25.8	10.5	1.3
New Zealand	4.7	11.6	21.7	26.4	22.3	10.7	2.7
Norway	6.0	13.6	24.8	28.9	19.0	6.4	1.1
Poland	1.3	7.7	22.5	33.1	24.5	9.1	1.7
Portugal	4.7	14.3	27.3	31.4	17.8	4.2	0.3
Slovak Republic	9.2	17.6	27.0	26.2	15.0	4.3	0.6
Slovenia	2.4	10.4	24.5	30.0	23.0	8.4	1.2
Spain	3.7	12.0	27.3	32.8	19.4	4.5	0.3
Sweden	7.3	15.0	26.2	28.0	17.2	5.6	0.7
Switzerland	3.0	9.8	22.8	31.3	23.7	8.3	1.0
Turkey	4.4	21.9	35.4	25.1	11.3	1.8	0.0
United Kingdom	4.3	10.7	22.4	28.4	23.0	9.3	1.8
United States	4.2	14.0	26.7	28.9	18.8	6.3	1.1
OECD average	4.8	13.0	24.5	28.8	20.5	7.2	1.1
Albania	23.5	29.6	28.5	14.4	3.6	0.4	0.0
Argentina	19.8	31.0	31.1	14.8	3.0	0.2	0.0
Brazil	18.6	35.1	30.7	12.5	2.8	0.3	0.0
Bulgaria	14.4	22.5	26.3	22.5	11.2	2.8	0.3
Colombia	19.8	36.3	30.8	11.0	1.9	0.1	0.0
Costa Rica	8.6	30.7	39.2	17.8	3.4	0.2	0.0
Croatia	3.2	14.0	29.1	31.4	17.6	4.3	0.3
Cyprus[5]	14.4	23.7	30.3	21.3	8.4	1.8	0.2
Hong Kong-China	1.2	4.4	13.0	29.8	34.9	14.9	1.8
Indonesia	24.7	41.9	26.3	6.5	0.6	0.0	0.0
Jordan	18.2	31.4	32.2	15.0	3.0	0.2	0.0
Kazakhstan	11.3	30.7	36.8	17.8	3.3	0.2	0.0
Latvia	1.8	10.5	28.2	35.1	20.0	4.0	0.3
Liechtenstein	0.8	9.6	22.0	30.8	26.7	9.1	1.0
Lithuania	3.4	12.7	27.6	32.9	18.3	4.7	0.4
Macao-China	1.4	7.4	22.2	36.2	26.2	6.2	0.4
Malaysia	14.5	31.0	33.9	16.5	3.7	0.3	0.0
Montenegro	18.7	32.0	29.7	15.4	3.8	0.4	0.0
Peru	31.5	37.0	23.5	7.0	1.0	0.0	0.0
Qatar	34.6	28.0	19.6	11.2	5.1	1.3	0.1
Romania	8.7	28.7	34.6	21.0	6.2	0.9	0.0
Russian Federation	3.6	15.1	30.1	31.2	15.7	3.9	0.3
Serbia	10.3	24.7	32.4	22.8	8.1	1.6	0.1
Shanghai-China	0.3	2.4	10.0	24.6	35.5	23.0	4.2
Singapore	2.2	7.4	16.7	24.0	27.0	16.9	5.8
Chinese Taipei	1.6	8.2	20.8	33.7	27.3	7.8	0.6
Thailand	7.0	26.6	37.5	21.6	6.4	0.9	0.1
Tunisia	21.3	34.0	31.1	11.7	1.8	0.1	0.0
United Arab Emirates	11.3	23.8	29.9	22.3	10.1	2.3	0.3
Uruguay	19.7	27.2	29.3	17.1	5.6	1.0	0.0
Viet Nam	0.9	5.8	20.7	37.5	27.0	7.1	1.0

Source: OECD, 2014.

[5] See footnote 2.

Figure 15. PISA 2012 science proficiency levels by country



The main points that can be observed from these figures and table are:

- The figure shows that within the OECD there are three countries which deviate from the pattern of average percentages for each proficiency level. These countries are Chile, Mexico and Turkey which have the smallest percentages of students in levels 5 and 6, but have substantial percentages of students in the lower proficiency levels. In Chile and Mexico there are over 30% of the students who are below level 2.

- In a manner similar to the results in reading and mathematics, in the partner countries/regions there is a much greater diversity in the pattern of percentages of students in the various proficiency levels. This diversity reflects large differences in economic status, resources committed to education, the nature of the educational programmes and cultural differences.

- Among the partner countries/region, there is a number of countries having over 60% of their students not able to reach level 2. These include Albania, Argentina, Brazil, Colombia, Indonesia, Jordan, Peru, Qatar and Tunisia.

- Among the partner countries/region, there is a number of countries with a notable percentage of students who are below the lowest described level - level 1. In can be seen in Figure 5 that this level is described as:

At level 1, students have such limited scientific knowledge that it can only be applied to a few familiar situations. They can present scientific explanations that are obvious and follow explicitly from given evidence.

- Peru (32%) and Qatar (35%) are the countries with the highest percentage of students who are below level 1, but there are three countries where more than 20% of the students do not reach level 1 - Albania, Indonesia and Tunisia. There are also six countries where there are between 18% and 20% of the students who do not reach level 1 – Argentina, Brazil, Colombia, Jordan, Montenegro and Uruguay. For these 11 countries PISA is unable to describe the skills that these students have - it is not that the students do not possess skills, it is just that the PISA survey is not designed to assess and describe them. Without this information it is extremely difficult for education systems to plan remediation programmes to help these students.

  Observation 17: When comparing reading, mathematics and science it is the last two which have the largest percentage of students below a described proficiency level – this is partly due to the fact that the described level 1 for reading was extended and divided into two sub-levels.

Conclusion

Across the three domains it can be seen that there is a significant percentage of students for whom PISA is unable to describe their proficiency in any way. This is more evident in mathematics and science than it is in reading. This result is probably a consequence of the expansion, in reading, of level 1 into two described proficiency levels – 1a and 1b.

Generally, for PISA to be more meaningful and more useful for nearly one third of the partner countries there needs to be a way to describe the proficiency of their lowest performing students. For countries currently not participating in PISA and in the process of considering participating, the lack of proficiency level description at the lower end may be acting as a deterrent.

Approaches to overcome this situation will be to:

- Consider how the existing items can be better used to describe the capacity of the students successfully completing them.
- Create new items that are directed at the lower end of the proficiency spectrum which are designed to elicit lower performing students' skills.

While creating new items of lower difficulty could be achieved, for example, by breaking down the assessed knowledge into smaller parts the challenge may be to keep the original PISA philosophy of assessing a student's preparedness for future life by measuring how they apply their existing knowledge.

For PISA 2009, an optional assessment, known as Reading Components was offered to countries. The aim of this was to break down reading tasks into smaller parts and assess the students' knowledge of those parts. Only two countries took part in the assessment, but it provided the opportunity for countries to assess something more fundamental to reading. This approach could one way for the PISA for Development project to proceed.

Other assessments taking place around the globe at a more fundamental level than PISA include the Early Grade Reading Assessment (EGRA) which assesses students at the early stages of their education. Important for consideration also are the ASER and UWEZO assessments which take place in India and east Africa respectively - these assessments are significant because they occur at home.

REVIEW OF SCALING MODELS

The item scaling model applied to PISA 2000-2012 is a generalised form of the Rasch model known as the multidimensional random coefficients Rasch model. The model is described by Adams and his colleagues, (Adams, Wilson and Wang, 1997; Adams and Wu, 2007; Adams, Wu and Carstensen, 2007) and has been implemented for the PISA data using the ACER *ConQuest*® software (Adams, Wu and Wilson, 2012).[6]

The Rasch model, in its general form was chosen for PISA for a number of reasons:

- Of all available item response theory models, it provides the strictest assessment of psychometric validity.
- It supports the construction and validation of meaningful described proficiency scales. Described proficiency scales are taken as a requirement for the useful reporting of PISA performance data.
- It has been widely generalised to deal with the range of analytic requirements of PISA. The Rasch family as implemented in ACER ConQuest can be, and is, used to explore and control for coder effects and item position (booklet) effects.
- Further, the model can be routinely applied in contexts that require multidimensional scaling.
- It supports equating tests for the purposes of maintaining and monitoring the validity of trends.

Alternative scaling models

As a part of its quality assurance oriented research programme, ACER undertook a number of comparisons of PISA's current Rasch model-based scaling with alternatives. The alternatives that have been researched include the use of a two-parameter logistic model and the use of unit-level scaling (see Macaskill, 2008).

As part of our research programme ACER re-scaled the PISA 2003 data with a two-parameter model and compared the outcomes with that obtained from Rasch model-based scaling. The key outcomes of that research programme were that:

Headline results, such as country means and variances, gender differences, country ranks and the like are identical under the Rasch and two-parameter logistic model scaling.

- Item-by-country interactions are more frequent under the two-parameter logistic model scaling than under the Rasch model scaling. That is, two-parameter model item parameter estimates vary more across countries than do Rasch model parameter estimates.

[6] The scaling approach that is planned for PISA 2015 is not known to the authors.

- With the exception of scoring information concerning partial credit items, the diagnostic feedback concerning the psychometric quality of items was the same for the two approaches.

- The partial credit items that are used in PISA may not be as discriminating as their Rasch model-based scoring assumes. In other words, too much weight may be being given to partial credit items.

There is no doubt that a range of modelling alternatives could be undertaken to improve the fit of the scaling models to the PISA data. And, indeed as we have discussed earlier there is compelling evidence that the misfit of the data to the PISA scaling models is greatest for developing non-Indo European language groups and non-western cultures. To reinforce this point Figures 16 to 19 are an (almost) random selection of comparisons between modelled and empirical item characteristic curves for PISA 2009 Tamil Nadu data. In each figure the smooth line is the modelled predicted proportion of correct responses for students at each proficiency based upon the PISA international parameter estimates. The dotted line is the actual proportion of successful responses for students at that proficiency level in Tamil Nadu. Discrepancy between the two lines indicates misfit in terms of an item-by-country interaction

Figure 16. Comparison of international modelled curve and empirical data for science item S521Q06 in Tamil Nadu

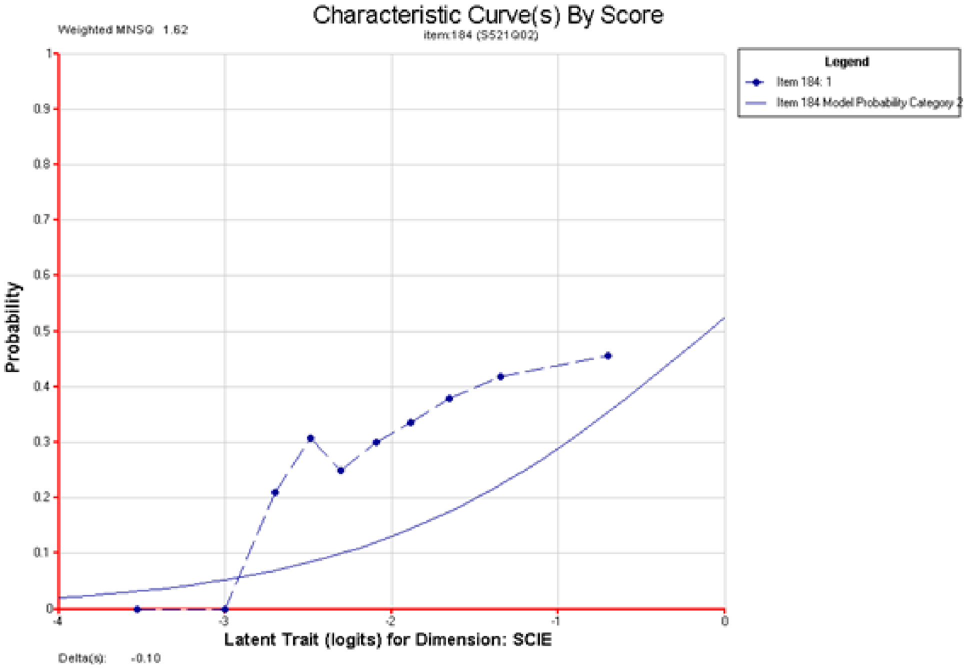


Figure 16 shows a Science item that was easier for students in Tamil Nadu than would be expected on the basis of their performance on all items. For example, for students with an estimated proficiency of around –1.0 logits the expectation internationally is that about 30% will be expected to respond correctly, whereas in Tamil Nadu about 45% of students at that level of proficiency responded correctly.

Figure 17 shows a Reading item for which the discrimination is quite different in Tamil Nadu than was modelled by the international parameters. At very low levels of proficiency very few if any Tamil Nadu students provided correct responses, whereas international 20-30% of students at the very lowest levels of proficiency where successful. At higher levels of proficiency, around –1.0 logits the international and Tamil Nadu proportions of correct response are quite similar.

Figure 17. Comparison of international modelled curve and empirical data for mathematics item R403Q03 in Tamil Nadu

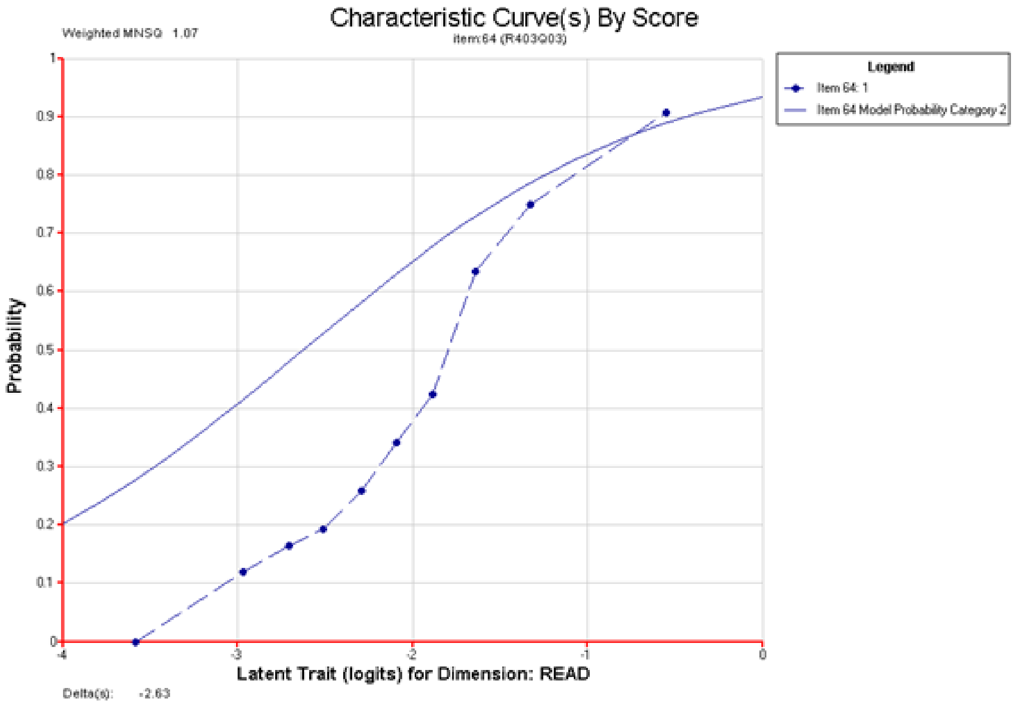


Figure 18 and Figure 19 are examples of Mathematics and Reading items, respectively that show possible evidence of guessing. For 2000-2012 the PISA scaling models have not taken the impact of guessing into consideration. The Mathematics item in Figure 18 is both easier overall than expected for Tamil Nadu, and further students at the lowest levels of proficiency are responding correctly at a much better than would be expected from the international model and parameter estimates. In contrast, the Reading item in Figure 19 is harder overall than expected for Tamil Nadu, yet students at the lowest levels of proficiency are responding correctly at a rate that matches expectation from the international model and parameter estimates. In each case the flattening of the empirical item characteristic curves at the low end suggests students are guessing and are provided correct responses at rate equal to chance.

There are two issues for PISA for development to consider in response to this evidence of misfit in the PISA scaling model in developing countries. The first is the superficial concern of relying on a model that does not fit. Should PISA for Development pursue alternative scaling models that admit allow features such as:

- varying discrimination across items;
- dependencies between items clustered in units;
- guessing; and
- parameter variation (including difficulty) across countries.

Observation 18: The fit of the developing country data to the PISA model is not good and scaling modifications to address some of the deviations should be explored. Changes to the model would however have wider implications for PISA, including a need to rescale previously collected data.

Figure 18. Comparison of international modelled curves and empirical data for mathematics item M564Q01 in Tamil Nadu

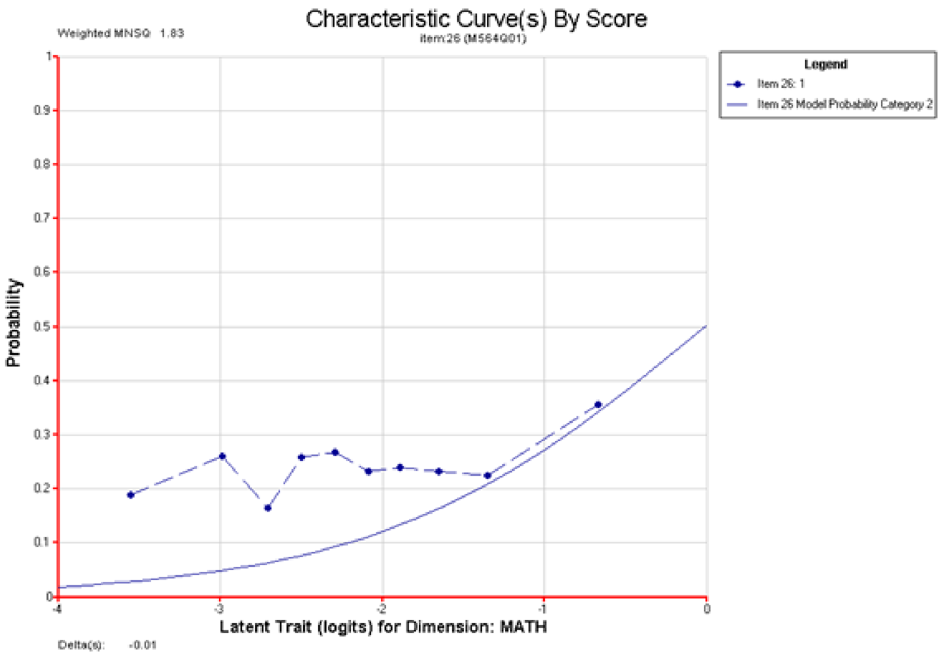
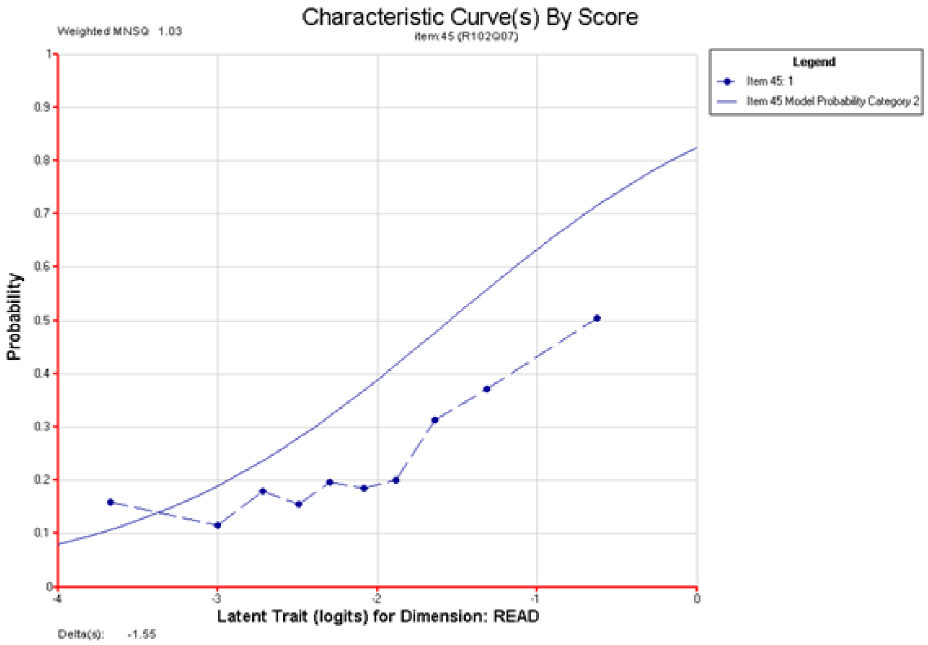


Figure 19. Comparison of international modelled curve and empirical data for reading item R102Q07 in Tamil Nadu



While modelling alternatives that deal with each of these issues do not present any technically insurmountable challenges they will not address the second and more profound issue of what is the implication of the misfit to the current models on the validity of PISA for Development results.

At the core of PISA are the *learning metrics* which are used to describe dimensions of educational progression. Each metric is depicted as a line with numerical gradations that quantify how much of the measured variable (*e.g.* reading) is present. Locations along this metric can be described by numerical scores or substantively (*i.e.* in terms of student skills, understanding and competencies). If the test performances of students who respond to different tests, who come from different participating countries and who respond to the tests at different times can be mapped on to common agreed learning metrics then valid international comparisons and trend reporting can be undertaken.

The value of these metrics is that they provide a substantive frame of reference that is stable and gives meaning to the otherwise uninterpretable test scores. Without the metrics we do not know if differences between test scores are due to differences between the items (*e.g.* over time or over alternative booklets) or differences between the students.

The concern with the item-by- country interactions that are observed here for developing countries is that a PISA-like learning metric cannot be constructed using the available secure items. This cannot be addressed with an alternative model since the core observation is that the pattern of item difficulties for the current PISA items varies widely across countries and it is hard to reconcile this with a common learning metric that can be used internationally to allow comparisons.

With access to a much larger pool of items it might be possible to revisit the issue of defining the PISA learning metrics on the basis of a common core of items that do behave in a (tolerably) consistent fashion across countries. This is not however possible here because the pool of items with appropriate difficulty is very small and further test development is not envisaged.

> Observation 19: The use of learning metrics to describe dimensions of educational progression is at the core of the PISA reporting methodology. This approach to reporting and construct validation requires a consistency across countries in item behaviour than is not apparent for developing countries.

OPTIONS

This paper has described the technical issues associated with the implementation of a targeted test for developing countries based on existing PISA items. There will be a number of different options which need to be considered when the programme is carried out.

Item selection

The paper has established clearly that it is indeed possible to design a test using the existing PISA items. This paper suggests that more difficult items could be excluded to arrive at a test better targeted at students of lower capacity. If countries agree to this option, then they will have a test which will give more precise information about their students' capacity allowing better measures of sub-populations to be comparisons for example because of a smaller standard error. Along with a more targeted context questionnaire this will lead to better analysis of the factors which are associated with student performance.

At the same time, however, because there will be no new items included, the existing items can only give the same descriptions of student capacity as in the regular PISA surveys. The result of this is that there may still be a significant proportion of students who fall below the lowest described proficiency level, so there will be no expansion of the knowledge gained about what students can and cannot do for those students.

The option of including only targeted items can suggest that there will be little or no opportunity for higher calibre students to demonstrate their skills. However this is not necessarily a concern because a sufficiently accurate measure of the students performing at a high level can be predicted with suitable modelling.

Sampling

Countries will need to consider options relating to the population of students who take the assessment. To maintain full comparability with the regular PISA survey the population must be 15-year-olds in educational institutions at grade 7 and above. To be consistent with the PISA philosophy of measuring preparedness for the future, countries may want to include 15-year-olds who are in grade 6 or below as it is possible those students may be leaving school within the near future. At the same time, countries may want to know how well the education system is preparing all students for the future especially at the end of middle secondary education and will want to include students in those grades who are older than 15. At the same time the test will be easier to organise within the school if all students within a grade are participating.

Countries may find also find it is easier to opt for a census approach, where all students of given ages are included. This has the advantage of better student engagement, because they haven't been singled out for participation, but has the disadvantage of being more costly in terms of test preparation/printing. test administration and data entry.

Test administration

With regard to test administration, the regular PISA survey employs independent administrators which go from school to school to carry out the test. Some countries may find it difficult to locate suitably qualified personnel for this task and will opt to use either ministry staff, student teachers or staff located at the school. Each of these options has its strengths and weaknesses. The employment of ministry staff may be convenient but takes them away from their regular tasks. The use of student teachers can be cost-effective and provide them with good experience, but they may have difficulty in coping with the number of students doing the test. The employment of local school staff is a model used in many assessments - this can be done in PISA, but there is a requirement that the staff member is not a teacher for any of the students in reading, mathematics or science.

Survey timing

Countries will need to consider options regarding the time that the test is given in the academic year – in the regular PISA survey testing does not take place in the first part of the year as this is seen as being a disadvantage for the students and, at the same time, the organisation of the test in the school takes some time.

In the regular PISA survey, the assessment is carried out every three years. An advantage for the countries undertaking the PISA for Development programme is that they are not tied into that level of frequency and may choose to do the assessment every four or five years. An advantage of this is that the costs are spread over a longer time period making participation more affordable.

Another significant advantage of spreading the frequency of testing over a longer time is that historically, the results of educational policy interventions take many years to occur and a longer time between testing may give more information than measuring every three years. The disadvantage of spreading the testing over a longer time is that the political cycle tends to be shorter and there is a need to be able to demonstrate that improvements in education have been made.

REFERENCES

Adams, R.J., A. Berezner and M. Jakubowski (2010), "Analysis of PISA 2006 preferred items ranking using the percentage correct method", *OECD Education Working Paper 46*, OECD Publishing.

Adams, R.J., M.R. Wilson and W. Wang (1997), "The multidimensional random coefficients multinomial logit model", *Applied Psychological Measurement*, Vol. 21, pp. 1-24.

Adams, R.J. and M.L. Wu (2007), "The mixed-coefficient multinomial logit model: A generalized form of the Rasch model", in M.v. Davier and C.H. Carstensen (eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications*, pp. 57-76, Springer Verlag.

Adams, R.J. and M.L. Wu (eds.) (2002), *PISA 2000 Technical Report,* OECD Publishing.

Adams, R.J., M.L. Wu and C.H. Carstensen (2007), "Application of multivariate Rasch models in international large scale educational assessment", in M.v. Davier and C.H. Carstensen (eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications*, pp. 271-280, Springer Verlag.

Adams, R.J., M.L. Wu and M.R. Wilson (2012), *ACER ConQuest Version 3: Generalised item response modelling software* [computer programme], Camberwell: Australian Council for Educational Research.

Grisay, A., J.H.A.L. de Jong, E. Gebhardt, A. Berezner and B. Halleux-Monseur (2007), "Translation Equivalence across PISA Countries", *Journal of Applied Measurement*, Vol. 8, No. 3, pp. 249-266.

Mackaskill, G. (2008), "Alternative Scaling Models and Dependencies", paper delivered at September 2008 Technical Advisory Group Meeting (TAG(0809)6a).

Mazzeo, J., E. Kulick, B. Tay-Lim and M. Perie (2006), *Technical Report for the 2000 Market-Basket Study in Mathematics*, (ETS-NAEP report #06-T01), Princeton, NJ: Educational Testing Service.

Mendelovits, J. (forthcoming), "Art and Science: Test Development in Large-Scale Educational Assessments", in J. Cresswell, P. Lietz, K. Rust and R. Adams (eds.), *The Implementation of Large Scale Educational Assessments*, Wiley, New York.

OECD (2014), *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*, PISA, OECD Publishing, http://dx.doi.org/10.1787/9789264208780-en.

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

OECD (2011), *Technical Report for the OECD Programme for International Student Assessment 2009*, OECD Publishing, http://dx.doi.org/10.1787/9789264167872-en.

OECD (2008), *Technical Report for the OECD Programme for International Student Assessment 2006*, OECD Publishing, http://dx.doi.org/10.1787/9789264048096-en.

OECD (2005), *Technical Report for the OECD Programme for International Student Assessment 2003*, OECD Publishing, http://dx.doi.org/10.1787/9789264010543-en.