Chapter 2: TEST DESIGN AND TEST DEVELOPMENT

INTRODUCTION

This chapter describes the assessment design for PISA for Development (PISA-D) as well as the processes used by the PISA-D contractor, Educational Testing Service (ETS), to select the cognitive instruments for the project. For PISA-D, under the guidance of the OECD and its partners, the decision was taken to offer a paper-based delivery survey that included the three core cognitive domains, with the aim to enhance the understanding of knowledge, skills, and contextual factors for students from a range of participating economies.

PISA-D assessment instruments were developed with the goal of providing reliable, valid, and comparable information from students in a wide range of low- and middle-income countries while ensuring that results are linked to the main PISA assessment. This design relied on the administration of paper-based assessment materials for 15-year-old students in grades 7 and above, as well as context (background) questionnaires for school administrators, teachers, and the person(s) most knowledgeable about the student (i.e., parent or guardian).

The development of the cognitive assessment was based on the following assumptions:

- a compulsory assessment of Reading, Mathematics, and Science, with equal weights for each of the three domains (i.e., no major/minor domain distinction as is made in PISA)
- paper-based cognitive instruments linked to PISA. This meant that a majority of items were selected from previous cycles of PISA but complemented with existing materials from surveys including PISA for Schools, the Programme for the International Assessment of Adult Competencies (PIAAC), the World Bank's Skills Towards Employability and Productivity (STEP) assessment, and the Literacy Assessment and Monitoring Program (LAMP)
- no new cognitive items
- all available items to be reviewed and selected to meet the measurement goals of PISA-D

The PISA 2015 trend item pool was considered the primary source of the PISA-D items. In order to obtain an accurate link to the PISA scales, item selection from the PISA 2015 trend item pool comprised at least half of the Main Survey assessment item pool. Based on the goal of PISA-D to provide enhanced coverage at the lower end of the three scales, the number of items representing Level 2 or below comprised approximately 60% of the PISA-D item pool, with items from the other existing international surveys noted above used to supplement them. Items from the other international surveys were selected because they map to aspects of the PISA frameworks and reflect the goals of the study. Items were selected with careful consideration of

the following criteria:

- maintaining intact units to the extent possible (sets of items with a common stimulus)
- ensuring adequate coverage of the key framework aspects
- an awareness of the cultural appropriateness of the contexts of the item stimuli
- an awareness of the amount of reading required for Mathematics and Science items

PISA-D INTEGRATED DESIGN

Goals and domain coverage

The assessment design for PISA-D was established with a total testing time for measuring the three domains—Reading, Math, and Science—of two hours for each student in both the Field Trial and the Main Survey. This timing is consistent with the timing for the main PISA assessment. The domain coverage specified in the design was intended to extend the range of information that PISA would provide to policy makers concerning the distribution of skills in their student populations. In summary, PISA-D was designed to provide participating countries with the following information:

- population distributions in Reading, Mathematics, and Science that reflect the PISA-D frameworks, as well as link to the most recent PISA core domain frameworks and scales reflected in the paper-based assessment; and
- pairwise covariance estimates among each of the three cognitive domains.

Table 2.1 shows the number of 30-minute clusters included in the PISA-D Field Trial and Main Survey. In order to meet the goals and domain coverage assumed in this design, each cluster was assembled from a combination of intact units of items from PISA 2015 and from existing surveys, and the items within each cluster represented a range of key framework aspects, item types, and item difficulties.

Table 2.1 **Cognitive domain coverage for PISA-D**

Cognitive Domain	Field Trial		Main Survey	
Reading	**5** 30-minutes clusters		**4** 30-minutes clusters	
	60% trend materials from PISA 2015	40% new materials from existing surveys	50-60% trend materials from PISA 2015	40-50% new materials from existing surveys
Mathematics	**5** 30-minutes clusters		**4** 30-minutes clusters	
	60% trend materials from PISA 2015	40% new materials from existing surveys	50-60% trend materials from PISA 2015	40-50% new materials from existing surveys
Science	**5** 30-minutes clusters		**4** 30-minutes clusters	
	60% trend materials from PISA 2015	40% new materials from existing surveys	50-60% trend materials from PISA 2015	40-50% new materials from existing surveys

Overview of the Field Trial Assessment Design

A Field Trial is an essential element in all surveys and is designed to yield information crucial for testing both instrumentation and survey operations. Data collection during the Field Trial was used to inform and refine final instruments and all procedures associated with the conduct for the Main Survey.

More specifically, the PISA for Development Field Trial was designed to meet the following key goals:

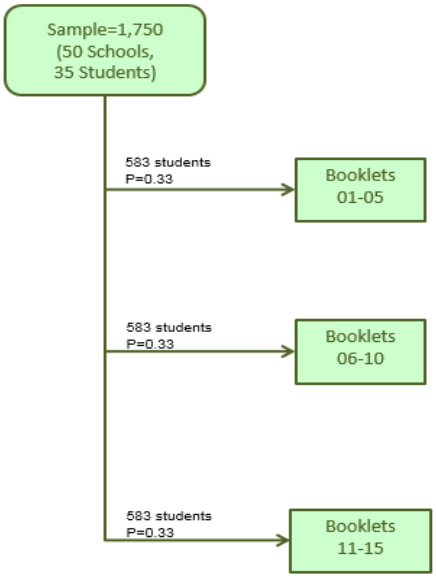
1. Operational goals: One of the purposes of the Field Trial was to evaluate the survey operation procedures. This included an examination of the efficiency and accuracy of data collection procedures, response rates for various subpopulations of interest, efficiency and accuracy of data processing (including recoding), and data submission.
2. Instrumentation goals: In addition to the examination of quality control measures on survey operations, the Field Trial also provided measures of the quality of the survey instruments, including the adequacy of scoring procedures, translation and adaptation quality, and scaling and analytic procedures.
3. Scaling and psychometric item characteristics: In order to support the comparability of inferences of PISA-D results across countries, including trend results with previous cycles of PISA, the equivalency of the psychometric characteristics of the items needed to be established. The PISA-D Field Trial data were used to examine the psychometric characteristics of the items and scales, and to evaluate the equivalence of item parameters with respect to trend items that provide a connection to prior PISA cycles. In addition, the Field Trial data provided initial data on the functioning of items that came

from other surveys and their appropriateness to the PISA-D population. These data were used to estimate preliminary item response theory (IRT) item parameters that served as a basis for selecting items from this additional item pool and constructing booklets for the Main Survey. The issues around equivalency and item selection were mainly addressed via IRT scaling using an innovative approach combining the Rasch model and the two-parameter logistic model as well as item analysis based on Classical Test Theory.

The Field Trial design, shown in Figure 2.1, required a reduced sample size that was based on a sample of 50 schools, with 35 students selected from each school for a total sample of 1,750 students per participating country. The design was based on five 30-minute clusters of items from each of the three domains. Within these clusters, approximately 60% of the items were expected to provide trend information from PISA 2015. The position of clusters for each domain was balanced and the assignment of a form to students followed an equal probability design. These clusters were combined and assembled into 15 booklets; each booklet measured two domains in order to provide covariance information. Each student received 60 minutes of assessment items, on average, in each of 2 domains.

■ Figure 2.1 ■

PISA-D Field Trial Assessment Design



Booklet		Cluster 1	Cluster 2		Cluster 3	Cluster 4
1	RC-Block1	R1	R2		S1	S2
2		S2	S3	RC-Block5	R2	R3
3	RC-Block3	R3	R4		S3	S4
4		S4	S5	RC-Block2	R4	R5
5	RC-Block5	R5	R1		S5	S1
6		S1	S2		M1	M2
7		M2	M3		S2	S3
8		S3	S4		M3	M4
9		M4	M5		S4	S5
10		S5	S1		M5	M1
11		M1	M2	RC-Block4	R1	R2
12	RC-Block2	R2	R3		M2	M3
13		M3	M4	RC-Block1	R3	R4
14	RC-Block4	R4	R5		M4	M5
15		M5	M1	RC-Block3	R5	R1

Where

- R1-R5 are Reading Literacy clusters
- RC-Block1-5 are Reading Components blocks
- M1-M5 are Mathematical Literacy clusters

- S1-S5 are Scientific Literacy clusters

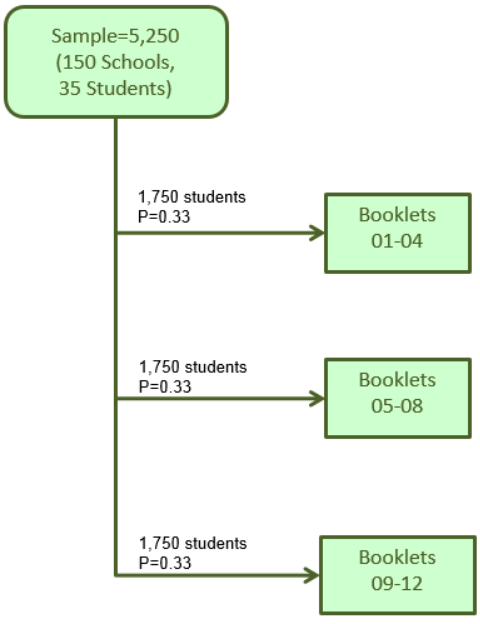
The findings of the Field Trial analyses are described in Chapter 9 of this report.

Overview of the Main Survey assessment design

The cognitive assessment design for PISA-D was planned so that the total testing time for measuring the three core domains of Reading, Mathematical, and Scientific Literacy remained at two hours for each student. An overview of the assessment design for the PISA-D Main Survey is provided in Figure 2.2. The cognitive assessment was to be administered to 35 students in each of 150 schools within each country. Further sampling requirements for this design are discussed in Chapter 4.

■ Figure 2.2 ■

Overview of the PISA-D Main Survey Assessment Design*



Booklet		Cluster 1	Cluster 2		Cluster 3	Cluster 4
1	RC-Block1	R1	R2		S1	S2
2		S2	S3	RC-Block2	R2	R3
3	RC-Block3	R3	R4		S3	S4
4		S4	S1	RC-Block4	R4	R1
5		S1	S2		M1	M2
6		M2	M3		S2	S3
7		S3	S4		M3	M4
8		M4	M1		S4	S1
9		M1	M2	RC-Block1	R1	R2
10	RC-Block2	R2	R3		M2	M3
11		M3	M4	RC-Block3	R3	R4
12	RC-Block4	R4	R1		M4	M1

Where

- R1-R4 are Reading Literacy clusters
- RC Block1-4 are Reading Components blocks
- M1-M4 are Mathematical Literacy clusters
- S1-S4 are Scientific Literacy clusters

THE PISA-D COGNITIVE FRAMEWORKS

For each PISA domain, an assessment framework is produced to guide instrument development and interpretation in accordance with the policy requirements of the PISA Governing Board. The

frameworks define the domains, describe the scope of the assessment, specify the structure of the test—including item format and the preferred distribution of items according to important framework variables—and outline the possibilities for reporting results. For PISA-D, Subject Matter Expert Groups (SMEGs) were convened by Pearson to review the existing PISA frameworks and provide suggestions for refinement of the descriptions of the performance of respondents who perform below level 2 of each of the cognitive scales. The SMEGs also reviewed the distributions of items across framework categories in PISA 2015 and made alternative recommendations, as appropriate, for PISA for Development. The expert groups' reviews and updates were based on the PISA 2012 and 2015 assessment frameworks.

PISA-D ITEM SELECTION

Item selection for the PISA-D Field Trial commenced in mid-2015. As the contractor for item selection, ETS was responsible for working with the subject matter experts in all domains to identify a suitable item pool from the available items—trend items from PISA 2015 and items from other available sources—based on the construct priorities established in the PISA-D frameworks and the requirements set forth in the integrated assessment design.

A proposed selection of trend items from the PISA 2015 item pool was shared with the SMEGs in July 2015 at the first meeting of the experts. Additionally, items developed for PISA for Schools, PIAAC, and LAMP were shared with the SMEGs for their input on their appropriateness with respect to the PISA frameworks, the target PISA population, and the targeted performance levels.

Response modes

Across all domains, PISA-D included items requiring one of two main response modes:

- Multiple choice, including either single-selection multiple choice or complex multiple choice (a table with statements and a number of yes/no or true/false options)
- Constructed response, including numeric and text entries that were coded either automatically or required human coding

PISA 2015 trend items

All available existing items (e.g., items that had not been released in previous cycles) from the PISA 2015 trend pool were initially considered for inclusion in the PISA-D cognitive assessment. Based on the assessment goals and selection criteria, an initial set of 54 Reading items, 45 Mathematics items, and 48 Science items was selected by the SMEG and recommended for the Field Trial.

New items selected from existing surveys

To meet the goals for PISA-D, a larger proportion of items measuring the lower end of the scale were needed than were available in the PISA trend pool. To supplement the PISA trend items, consideration was given to items from existing surveys. For Reading, available items from PIAAC, LAMP, STEP, and PISA for Schools were reviewed and selected. Included among the Reading items was a set of Reading Components, which include shorter sentence processing and passage comprehension items used in STEP, PIAAC, and PISA. Some of these existing Reading

Components items were modified with additional response options. In addition, some Reading stimuli and items from existing surveys were adapted to better meet the needs for the assessment design. For Mathematics, available items from PIAAC and PISA for Schools were selected. Modifications were made to some items to simplify the reading load, and to some coding guides to introduce partial-credit coding to better differentiate student responses and understand characteristics of mathematics proficiency at the lower level of the PISA scale. For Science, available items from PISA for Schools were selected, along with some new PISA items developed for the computer-based assessment of Science in 2015. As with the items for Reading and Mathematics, some items were adapted from the original source versions to simplify the stimuli and tailor the items to meet the goals for PISA for Development. When items were adapted from their original version, they were no longer treated as trend items.

National reviews and selection of Field Trial items

A second stage of review and recommendations for selection involved the participating PISA-D countries, where the National Centres' staff had the opportunity to review and provide feedback on units recommended for the Field Trial by the SMEG and international test development team. Feedback was provided by all seven participating countries and included countries' evaluation of the appropriateness of the classifications of the items according to the framework aspects, the suitability of the context for 15-year-olds, and any additional comments about the appropriateness of the item for the assessment. Items were selected for inclusion in the Field Trial using an overall judgment based on country reviews, feedback from the expert group, and the distribution of items across the key categories as defined in the framework.

FIELD TRIAL

The PISA-D Field Trial data collection timeline began in September 2016 and extended through December 2016 with seven participating countries or economies across four language versions. Assessment materials were prepared and released based on the Field Trial testing dates for each country.

Preparation of Field Trial instruments

As part of the quality control procedures for PISA-D, ETS assumed responsibility for managing the process of assembling the paper-based versions of the cognitive instruments, preparing all paper booklets used in the Field Trial. Countries were responsible for adapting and/or translating all material and performing both linguistic and layout quality control checks for items. Where countries identified errors as a result of those checks, they were shared with the contractors who made any agreed-upon corrections.

The approved clusters were then assembled into the 15 Field Trial paper booklets by the contractors in a centralised fashion that ensured comparability of layout. As a final step, booklets were released to countries so that the sequence of clusters within forms could be confirmed and, once approved, print-ready versions were provided to National Centres. Once those had been corrected and their paper booklets assembled, they were asked to check and sign off on the final instruments.

Field Trial coding

Coding guides for all PISA-D items were compiled by ETS, in cooperation with cApStAn, based on the existing versions of the guides. The English and French source versions and Spanish base version of the coding guides were released in draft form prior to the coder training meeting in July 2016. Based on discussions at that meeting, the coding guides were finalised and updated source and base versions were released to countries in August 2016, prior to the beginning of the Field Trial data collection period.

Field Trial coder training

The international Field Trial coder training was held in July 2016 and focused on all domains and all items. The goals of the training included both having attendees develop an in-depth understanding of the coding process for each item so they would be prepared to train coders in their countries and reaching consensus about the coding rules to better ensure consistency of coding within and between countries and across cycles. Trainers reviewed the layout of the coding guides, general coding principles, common problems, and guidelines for applying special codes. Sample student responses were provided and attendees were required to code them. Where there were disagreements about coding for a particular item, those were discussed so that all attendees understood, and would be able to follow, the intent of the coding guides.

Field Trial coder queries

As was the case in PISA, ETS set up a coder query service for the PISA-D Field Trial. Countries were encouraged to send queries to the service so that a common adjudication process was consistently applied to all coder questions about constructed-response items. Queries were reviewed and responses provided by domain-specific teams.

In addition to responses to new queries, the queries report included the accumulated responses from previous cycles of PISA for all PISA 2015 trend items included in PISA-D. This helped foster consistent coding of trend items. The report was regularly updated as new queries were received and processed and National Centres were notified as updates were posted.

Field Trial outcomes

The PISA-D Field Trial was designed to yield information about the quantity and quality of data collected. This information was crucial for the selection and assembly of the Main Survey instruments and for refining survey procedures where necessary. More specifically, the goals of the Field Trial included collecting and analysing information regarding:

- the quantity of data and the impact, if any, that survey operations had on that data;
- the quality of the items; and
- the use of the data to establish reliable, valid, and comparable scales based on IRT models.

Details about the Field Trial analysis are discussed in Chapter 9.

MAIN SURVEY

The PISA-D Main Survey began in September 2017 and ended in late December 2017. In preparation for the Main Survey, countries reviewed items based on their performance in the Field Trial and were asked to identify any serious errors still in need of correction. The international contractors worked with countries to resolve any remaining issues and prepare the national instruments for the Main Survey.

National item review following the Field Trial

The item feedback process began in May 2017 and concluded in October 2017. The process involved countries reporting any linguistic or layout issues that were noted during the Field Trial, including errors to the coding guides. Following release of the Field Trial data, countries received item feedback forms that included flags for any items that had been identified as not fitting the international parameters. Flagged items were reviewed by national teams. Countries were asked to provide comments about these specific items where they could identify serious errors. Requests for corrections were reviewed by cApStAn and, where approved, implemented.

Item selection

The initial selection of items recommended for the Main Survey was made by the test development team based on item statistics from the Field Trial, country comments, coverage of the domain as specified in the framework, item format, and the assessment design.

In April 2017, the SMEGs met to review the Field Trial results and recommend the final item pool for the Main Survey. The experts reviewed the Field Trial data, which included summary item statistics, the item key or coding guide, framework classifications, and notes from the psychometricians. The experts reviewed a proposed set of items, which took into account the goals and constraints for item selection and took into account the data quality. As a result of their discussions, a small number of items were dropped from the recommended pool and replaced by alternates, and suggested changes were made to the order of units within each cluster based on considerations of content and item difficulties. For the Mathematics items, the experts identified some difficult open-ended items for which incorporating partial credit within the coding guides needed to be considered to enhance the understanding of the processes undertaken when solving mathematical problems.

Construct coverage

The set of items for the Main Survey was balanced in terms of construct representation to the extent possible, given the constraints of the assessment, based on the overall distributions recommended in the frameworks.

A total of 67 items were selected for Reading, with the distribution as shown in Table 2.2 below.

Table 2.2 **Reading item counts by framework category**

Process	Items	Percent	Framework Goal
Access and retrieve	22	33%	25-30%
Integrate and interpret	31	46%	45-55%
Reflect and evaluate	14	21%	15-25%
Situation			
Personal	22	33%	25-45%
Educational	21	31%	25-45%
Occupational	4	6%	15-25%
Public	20	30%	5-15%

A total of 64 items were selected for Mathematics, with the distribution as shown in Table 2.3 below.

Table 2.3 **Mathematics item counts by framework category**

Process	Items	Percent	Framework Goal
Formulate situations mathematically	13	21%	Approx. 25%
Employing mathematical concepts, facts, procedures	28	44%	Approx. 50%
Interpreting, applying, and evaluating mathematical outcomes	22	35%	Approx. 25%
Context			
Change and relationships	12	19%	25%
Space and shape	9	14%	25%
Quantity	27	43%	25%
Uncertainty and data	15	24%	25%

A total of 66 items were selected for Science, with the distribution as shown in Table 2.4 below.

Table 2.4 **Science item counts by framework category**

Competency	Items	Percent	Framework Goal
Explaining phenomena scientifically	35	53%	40-50%
Evaluate and design scientific enquiry	13	20%	20-30%
Interpreting data and evidence scientifically	18	27%	30-40%
Scientific Knowledge			
Content	22	64%	55-65%
Procedural	21	25%	20-30%
Epistemic	4	12%	10-20%

Main Survey coding

The process used for the Main Survey coding training was identical to that employed prior to the Field Trial. Full training was provided for all items across the domains. Special attention was given to the coding design for the Main Survey, which was more sophisticated than that for the Field Trial.

The coder query service was again used in the Main Survey as it had been in the Field Trial to assist countries in clarifying any uncertainty around the coding process or responses. Queries were reviewed and responses provided by domain-specific teams.

Review of Main Survey item analyses

The Main Survey data went through extensive analyses implemented through multi-step procedures to ensure the quality of the results. The first steps were implemented to evaluate the overall quality of the data submitted by countries looking at how well the assessment design and booklet assignment were reflected in the data as well as looking for the effects of any possible threats to data quality, such as scoring inconsistencies and other administration problems. These were followed by more specific analyses including item analysis; coding and treatment of missing data; item response theory scaling, including international item fit and item-by-country interactions; conditioning models; and generation of plausible values. These procedures are described in more detail in Chapters 9, 10, and 12. Finally, the outcomes of these analyses guided decisions around data products and treatment of items as described in detail in Chapter 19.

Reference

OECD (2018), *PISA for Development Assessment and Analytical Framework: Reading, Mathematics and Science,* OECD Publishing, Paris, *http://dx.doi.org/10.1787/9789264305274-en.*