Chapter 2: TEST DESIGN AND TEST DEVELOPMENT

INTRODUCTION

This chapter describes the assessment design for PISA for Development (PISA-D) Strand C as well as the processes used by the PISA-D contractor, Educational Testing Service (ETS), to select and prepare the cognitive assessment instruments for the project. As such, under the guidance of the OECD and its partners, the decision was taken to offer a computer-based household survey for out-of-school populations, including a Youth Interview and a cognitive test of Reading and Mathematics.

The scientific literacy domain was not included in the Strand C assessment due to practical considerations of total assessment time and the burden on individuals in a household survey. On one hand, the total test allowed a maximum of 50 minutes, which did not provide adequate time to include an assessment of three domains and meet the requirements of linking to the PISA scales. Therefore it became necessary to choose only two domains. In making the decision, it was taken into account that reading and mathematics literacy are considered foundational skills that are necessary for the development of scientific literacy skills. In addition, the target population was also taken into account. As science is the domain with the strongest link to school-based learning, this domain was the least appropriate for a group that, by definition, has been exposed to less formal schooling. Therefore, it was decided to include reading and mathematics as two domains assessed in Strand C.

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 could be reported on the main PISA assessment scale. Specifically, PISA-D Strand C is a household survey that aims to accurately describe the proficiency of 14- to 16-year-olds in each country, who are either out of school or in school at grades 6 or lower, on scales comparable to those administered in the in-school strand of PISA-D. Due to the unique nature of the target population, sampling and survey procedures for PISA-D Strand C were different from those for the school-based survey. This design relied on the administration of computer-based assessment materials, as well as a contextual (background) questionnaire for the youth and one for the person most knowledgeable about the youth (e.g. parent or guardian). These were administered by trained interviewers during household visits, using tablet computers rather than paper booklets. The PISA-D Strand C design also included a household observation questionnaire completed by the interviewer that asked questions about the location of the household, aspects of the neighbourhood, characteristics of the dwelling and general conditions of the assessment administration.

The development of the PISA-D Strand C cognitive assessment was based on the following assumptions:

- a compulsory assessment of Reading and Mathematics, with equal weights for each domain (i.e. no major/minor domain distinction as is made in PISA)
- computer-based cognitive instruments linked to PISA, using a subset of the items chosen for PISA-D Strand A
- the distribution of items selected for PISA-D Strand C would focus on the lower end of the difficulty scale.

Reading and Mathematical Literacy items for PISA-D Strand C were selected from the PISA-D Strand A items, which in turn were drawn primarily from the PISA 2015 trend item pool. This allowed for the establishment of a link between the reported results and the PISA scales. Based on the goal of PISA-D Strand C to provide enhanced coverage at the lowest end of the proficiency scales, the majority of items selected for PISA-D Strand C were items located at PISA's proficiency Level 2 or below. 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 items.

Items selected for PISA-D Strand C were adapted for a computer-based administration, as the survey was designed to be delivered only on tablet computers. Tablet-based administration also allowed for the implementation of an adaptive test design that routes respondents to a particular branch of the cognitive assessment based on the number of correct responses to an initial set of core questions. The Youth Interview for PISA-D Strand C included a subset of items from the Student Questionnaire administered in PISA-D Strand B, but with corresponding modifications to account for the interviewer-delivered instrument. Additional items were included in the Youth Interview to identify respondents' experience within the educational system and their work history. The development and content of the PISA-D Strand C questionnaires is described in Chapter 3.

PISA-D STRAND C INTEGRATED DESIGN

Goals and domain coverage

The cognitive assessment design for PISA-D Strand C was established with a total testing time for measuring the two domains—Reading and Mathematical Literacy—of about 45 minutes for each student in both the Field Trial and the Main Survey. Because the assessment was delivered individually, and at home, actual testing time per individual varied. 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 and proficiency in 14- to 16-year-olds who either are not in school or are at grade 6 or below. In summary, PISA-D Strand C was designed to provide participating countries with the following information:

- population distributions in Reading and Mathematics 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 (PBA)
- covariance estimates between the two assessment domains, mathematics and reading literacy.

Table 2.1 shows the number of items and clusters included in the PISA-D Strand C Field Trial and Main Survey. In order to meet the goals and domain coverage assumed in this design, each cluster was assembled from items from PISA-D Strand A, and the items within each cluster represented a range of key framework aspects, and item types.

Table 2.1 Cognitive domain coverage for PISA-D Strand C

Cognitive Domain	Field Trial	Main Survey
Reading Literacy	**36 items** (3 clusters of approx. 12 items in each cluster)	**22 items** 5 items in Core 17 items divided among 3 clusters (approx. 10 items per cluster)
Mathematical Literacy	**45 items** (3 clusters of approx. 15 items in each cluster)	**34 items** 5 items in Core 30 items divided among 3 clusters (approx. 10 items per cluster)
Reading Components (Sentence Processing and Paragraph Comprehension)	**66 items** (3 clusters with 32 common items and approx. 8 unique items in each cluster)	**51 items** 24 Sentence Processing items in each of 3 clusters 27 Paragraph Comprehension items divided among 3 clusters (approx. 9 items per cluster)

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 instrumentation, as well as sampling and survey operations. Data collection during the Field Trial was used to inform and refine final instruments and all other designs and procedures associated with the conduct for the Main Survey.

More specifically, the PISA-D Strand C Field Trial was designed to meet the following key goals:

1. Sampling and survey operations goals: One of the purposes of the Field Trial was to evaluate the sample design and selection, and to evaluate and practice the survey operation procedures. This included an examination of the sample design and selection procedures at various stages, 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 and comparability with Strands A and B, 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 Strand C population. These data were used to estimate preliminary item response theory (IRT) item parameters that served as a basis for selecting items for the Main Survey.
4. Computer-delivery platform and Case Management System: As PISA-D Strand C was implemented through the use of tablets, the Field Trial addressed the following tasks related to an interview-based study: i) test and evaluate the functioning of the cognitive assessment portion of the delivery platform, particularly response capturing and automatic scoring; ii) test and evaluate the functioning of the computer-assisted personal interviewing (CAPI) system, particularly the flow of questions and efficiency of the system in capturing information; iii) evaluate the accuracy of the interviewer's instructions; iv) test the operational effectiveness of the system during the interview; and v) test the system for assigning cases to interviewers, storing case files and managing reports at the national level.

The Field Trial design, shown in Table 2.2, required a sample size that was based on a total sample of 1 200 students per participating country. The design included items from each of the two cognitive assessment domains distributed across 18 different forms.

Table 2.2 PISA-D Strand C Field Trial assessment design

Forms	Part 1		Part 2	
1	RC-A	R1	R2	
2	RC-B	R2	R3	
3	RC-C	R3	R1	
4	RC-B	R1	R3	
5	RC-C	R2	R1	
6	RC-A	R3	R2	
7	RC-B	R1	M1	
8	RC-C	R2	M2	
9	RC-A	R3	M3	
10	M1		RC-A	R1
11	M2		RC-B	R2
12	M3		RC-C	R3
13	M1		M2	
14	M2		M3	

Forms	Part 1	Part 2
15	M3	M1
16	M1	M3
17	M2	M1
18	M3	M2

Where

- R1–R3 are Reading Literacy clusters
- RC-A, RC-B, and RC-C are Reading Components blocks
- M1–M3 are Mathematics Literacy clusters.

Each respondent was assigned to one of 18 forms. Each form consisted of two clusters of items. Reading clusters were always paired with a Reading Components block as shown in the design. There were 6 forms that included only Reading items, 6 forms included a combination of Reading and Mathematics items, and 6 forms included only Mathematics items.

The PISA-D Strand C Field Trial proved to be a challenge for participating countries to meet the required Field Trial sample sizes. The 1 200 required sample size was needed to evaluate the Field Trial goals noted above. In particular, the scaling and psychometric goal of 400 responses per country for the estimation of item parameters. Since by design, not every youth takes every assessment item, the recommended sample size of 1 200 respondents was necessary to ensure an adequate number of responses per item.

Of the six countries participating in the Field Trial, only two (Guatemala and Senegal) met the sample size requirements. Regardless, the sample sizes across all participating countries were insufficient for the Field Trial analyses to be performed as planned and be used for finalising the Main Survey instruments and design. Therefore, while some of the Field Trial goals were met, the goal of scaling and evaluating the psychometric characteristics of the items was not met and some of the analytical burden that was originally planned for the Field Trial became an obligation for the Main Survey. 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 Strand C was planned so that the total testing time for measuring the two core domains of Reading and Mathematical Literacy was approximately 45 minutes, on average, for each student. An overview of the assessment design for the PISA-D Strand C Main Survey is provided in Figure 2.2, and includes the following design characteristics:

- All instruments were delivered on a tablet that captured, stored and exported all data.
- The Youth Interview was administered to the respondent by an interviewer. Following this questionnaire, the tablet was passed to an eligible respondent to begin the

cognitive assessment.

- The Main Survey cognitive instrument included a 10-minute *Core Module* of basic reading and mathematics skills to ensure that respondents have an appropriate level of skills to proceed to the full assessment. An established minimum number of items answered correctly determine the next set of items to be presented to respondents in the second stage of the cognitive assessment. The second stage was designed to take no longer than 35 minutes to complete.
- Respondents who pass the Core Module were randomly assigned to one of the 12 forms measuring Reading and Mathematical Literacy (shown on the right-hand branch in Figure 2.2 below). The passing score for the Core Module was at least one item per domain answered correctly. All forms included items from both domains. Forms 1-6 were comprised of two Reading and a single Mathematical Literacy cluster. Forms 6-12 were comprised of two Mathematical literacy and a single Reading cluster. A cluster of Reading Components items was administered within each form prior to the first Reading cluster in the corresponding form.
- Respondents who failed the Core Module were directed to an assessment comprised of all Reading Components items. This was expected to take no longer than 15 minutes.

During the main survey, the cognitive assessment was to be administered to a minimum of 1 600 respondents within each country, split into at least 1 200 from a representative sample and 400 from a limited representative sample. Further sampling requirements for this design are presented in Chapter 4.

Figure 2.1 Overview of the PISA-D Main Survey assessment design

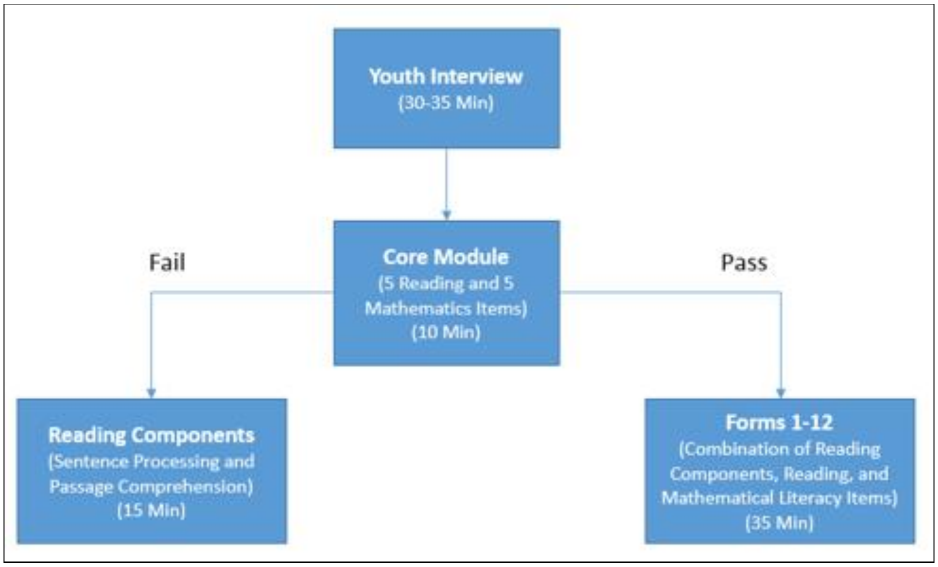
Youth Interview
(30-35 Min)
Core Module
(5 Reading and 5 Mathematics Items)
(10 Min)
Fail
Pass
Reading Components
(Sentence Processing and Passage Comprehension)
(15 Min)
Forms 1-12
(Combination of Reading Components, Reading, and Mathematical Literacy Items)
(35 Min)

Table 2.3 PISA-D Strand C Main Survey assessment design

Form	Part 1	Part 2	Part 3	Part 4
1	RC-A	R1	R2	M1
2	RC-B	R2	R3	M2
3	RC-C	R3	R1	M3
4	M1	RC-B	R1	R3
5	M2	RC-C	R2	R1
6	M3	RC-A	R3	R2
7	M1	M2	RC-B	R1
8	M2	M3	RC-C	R2
9	M3	M1	RC-A	R3
10	RC-A	R1	M1	M3
11	RC-B	R2	M2	M1
12	RC-C	R3	M3	M2

Where

- R1–R3 are Reading clusters
- RC-A, RC-B, and RC-C are Reading Components blocks
- M1–M3 are Mathematics clusters.

During the main survey, each respondent "passing" the core module would respond to one Reading Component block, one or two math clusters and one or two reading clusters.

THE PISA-D STRAND C 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 PISA's proficiency Level 2 in 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-D. The expert groups' reviews and updates were based on the PISA 2012 and 2015 assessment frameworks.

PISA-D STRAND C ITEM SELECTION

Items were selected for the PISA-D Strand C Field Trial from the Strand A Reading and Mathematics item pools. The selection focused on three main criteria: item difficulty, framework coverage and adaptability to tablet computer delivery. Items were selected at the unit level with the goal of keeping units intact. Units with most items representing the capacity of individuals at PISA's proficiency Level 2 and below were considered for the PISA-D Strand C instruments.

The full set of units selected represented the range of constructs and aspects of the framework. The integrated design specified the need for an adaptive design for the Main Survey based on a total correct score for the Core items. To this end, cognitive units with multiple-choice items or open-ended items for which scoring could be automated were preferred for the core.

The main response mode of single or multiple-choice selection involved the respondent using a finger to select a response option either by tapping on text within the stimulus[1] or by selecting a radio button next to a response from a limited set of options. Another common response mode was numeric entry. Numeric entry required a respondent to type a numeric response using number keys and a decimal point (either a period or comma as appropriate, depending on local conventions). In this response mode, the delivery system prevented the respondent from including any text. A third response mode involved the respondent using a finger to drag and drop a selection into a pre-defined area.

Previous experience with large-scale international assessments such as PISA and the Programme for the International Assessment of Adult Competencies (PIAAC) has shown that item parameters for paper-and-pencil items were not significantly impacted when adapted for computer administration. The process of adapting paper-and-pencil items to computer administered and scored items carefully considered the type of modifications of the response mode to ensure the demands of computer skills by the respondent were limited. An interactive tutorial was designed to be administered prior to responding to the assessment items. This tutorial helped establish familiarity with the response mode and included simple practice exercises to be completed before the assessment began.

FIELD TRIAL

The PISA-D Strand C Field Trial data collection timeline began in March 2017 and extended through August 2017 with six participating countries, across three language versions. Assessment materials were prepared and released based on the Field Trial testing dates for each country.

Preparation of Field Trial instruments

A master international computer-based version of the Field Trial units was finalised in February 2017. Since translation of the items was initially completed for most PISA-D Strand C countries as part of the translation process for PISA-D Strand A, the PISA-D contractors centrally implemented any minor changes to items, which included typos or grammatical errors from the PISA-D Strand A Main Survey as well as global changes to directions or instructions due to the adaptations to response modes between the paper-based and computer-based versions (e.g. changing “circle” to “tap on” in the directions). Further details of the translation and adaptation process are described in Chapter 3 of this Technical Report.

[1] PISA tests employ a range of stimuli as the basis for items. The stimulus reflects the wide variety of text types and topics that 15 year-old readers encounter every day. These may include, but are not limited to, excerpts from novels, transcriptions of interviews, academic articles, short stories or fables, instruction manuals, reviews, blog posts, job descriptions, letters, web sites, wiki entries, catalog entries, charts and diagrams, brochures, advertisements, product descriptions, and newspaper or magazine articles.

Following the translation, adaptation and verification processes, checks of the layout of the items were completed to verify that the adaptations did not introduce any unexpected layout changes that would impact the display of the items. Additional checks were completed by the PISA-D contractors to verify that the automated scoring rules for the open-ended items were properly implemented. Once the master versions of the computer-based units were reviewed and finalised, they were assembled into the clusters for administration according to the assessment design and incorporated into the PISA-D Strand C delivery system.

Details about the Field Trial analysis and results are discussed in Chapter 9 of this Technical Report.

MAIN SURVEY

The PISA-D Strand C Main Survey data collection began in September 2018 and ended in late January 2019. In preparation for the Main Survey, PISA-D contractors reviewed the data available from the Field Trial and selected a reduced item pool for the Main Survey to meet the assessment goals while also respecting the burden of testing time on the respondents.

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, coverage of the domain as specified in the framework, item format and the assessment design.

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 22 items and 34 items were selected for Reading and Mathematics, respectively, with the corresponding distributions by framework categories shown below in Table 2.4 and Table 2.5.

Table 2.4 Reading item counts by framework category

Process	Items	Percent	Framework Goal
Access and retrieve	8	36%	25-30%
Integrate and interpret	12	55%	45-55%
Reflect and evaluate	2	9%	15-25%
Situation	**Items**	**Percent**	**Framework Goal**
Personal	5	23%	25-45%
Educational	10	45%	25-45%
Occupational	1	4%	15-25%
Public	6	27%	5-15%

Table 2.5 Mathematics item counts by framework category

Process	Items	Percent	Framework Goal
Formulate situations mathematically	7	21%	Approx. 25%
Employing mathematical concepts, facts, procedures	16	47%	Approx. 50%
Interpreting, applying, and evaluating mathematical outcomes	11	33%	Approx. 25%
Context	**Items**	**Percent**	**Framework Goal**
Change and relationships	6	17%	25%
Space and shape	14	41%	25%
Quantity	7	21%	25%
Uncertainty and data	7	21%	25%

RECOMMENDATIONS

The Main Survey data went through extensive analyses to examine and evaluate the quality of the results. The outcomes of the Main Survey data analyses, described in detail in Chapter 13, guided decisions around data products and treatment of items. Given that a large number of participants failed the Core assessment, the data suggest that further development of a test of basic reading and mathematics skills for out-of-school youth would benefit from new item development focused on those tasks measuring the lowest proficiency levels of the PISA scale (Levels 1a, 1b and 1c) and expanding below Level 1c to examine more basic literacy and numeracy skills.

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.