Chapter 2
Test Design and Test Development

INTRODUCTION

This chapter describes the assessment design for PISA 2018 as well as the processes used by the PISA Core A contractor, Educational Testing Service (ETS), and the international test development team to develop the tests for the 2018 cycle.

The tests for the 2018 cycle included the following:

- a reading test, the major domain in 2018
- a mathematics and a science test, the two minor domains
- a global competence test, the innovative domain for this cycle, and
- a financial literacy test, an international option for this cycle.

In the 2015 cycle, PISA moved from a primarily paper-based delivery survey that included optional computer-based modules, to a fully computer-delivered survey. A paper-based version of the assessment that included only trend units was developed for the small number of countries that did not implement the computer-delivered survey. The 2018 cycle retained this same option, using the same paper-based materials as the 2015 cycle. The computer-based delivery mode allows PISA to measure new and expanded aspects of the domain constructs. In reading, new material included multiple sources and digital reading formats to better represent the kinds of reading that are becoming more prevalent in the 21st century. Computer-based multistage adaptive testing was also adopted for the reading literacy domain to further improve measurement accuracy and efficiency, especially at the extremes of the proficiency scale. In financial literacy, some interactive tasks were created that allowed students to manipulate variables and observe effects of financial choices.

PISA 2018 INTEGRATED DESIGN

The goals for the integrated assessment design in PISA 2018 included:

- improving the measurement of trends over time across the three core PISA domains (reading, mathematics and science)
- minimising respondent burden, while maximising the range of information obtained for each domain assessed
- accurately describing the proficiencies of nationally representative samples of 15-year-olds in each country, including relevant subpopulations
- associating these proficiencies with a range of indicators in policy-relevant areas.

To meet these goals, the design for PISA 2018 was based on the design and methodological innovations first introduced in the 2015 cycle. In contrast to cycles prior to 2015 where scaling was focused on the cycle at hand and required a new scoring transformation each time, the methodology introduced in 2015 incorporated all available data for scaling and provided a scoring transformation applicable to 2015 as well as future cycles. It provided a solid base for linking across cycles and between paper- and computer-based administrations for all cognitive scales and facilitated the development and transition to a computer-based adaptive testing.

As a form of adaptive testing particularly well suited for PISA, multistage adaptive testing was adopted in 2018 for the reading literacy domain to reduce measurement error for heterogeneous populations without overburdening individual respondents. Taken together, these design and methodological innovations served to improve comparability across countries, stabilise parameter estimations and the measurement of trends, and improve the reliability of inferences formed from the data. In addition, as part of the design for 2018, ETS fully integrated the innovative domain of global competence into the assessment design together with the core domains of reading, mathematical and scientific literacy.

Minimising the distinction between major and minor domain coverage

Prior to 2015, the PISA design focused on keeping the number of students who responded to each item in both the major and minor domains relatively constant. As a result, as shown in Table 2.1 below, the number of items included in the minor domains was significantly less than the number of items in the major domain for that year (marked with an asterisk). Note, for example, that when reading was a minor domain in 2003 and 2006, it contained only about 20% of the items included when it was the major domain in 2000. Across the other two core domains and cycles, the number of items included when a domain was minor ranged from about 35% to no more than 57% of the total number included when the domain was major.

In contrast, under the assessment design for 2015, 103 items were used in the minor domain of reading — close to 80% of the items included when reading was last the major domain in 2009 — and there were 83 items in mathematics, or just over 75% of the items used when mathematics was the major domain in 2012. This increase in the number of items was maintained for the minor domains of scientific and mathematical literacy in the 2018 cycle.

Table 2.1: Number of PISA items by domain and across cycles in the main survey (numbers in red indicate the major domain in each cycle)

Altogether, the scaling approach implemented in PISA 2018 and the inclusion of a larger number of items in each minor domain helped stabilize and improve the measurement of trend by making the construct coverage for each minor domain comparable to that of a major domain. But consequently, the trade-off is a reduction of the number of student responses per item when the domain is a minor domain.

Under this approach for measuring trends, each domain goes through a “domain rotation”, or a nine-year period that begins with a new or revised framework and continues with the two subsequent cycles in which it becomes a minor domain and then concludes with becoming a major domain once again. The end of the 9-year domain cycle involves another revision of the framework to reflect the current thinking about assessment for the new data collection as a major domain. For example, the revised framework for reading as the major domain in 2018 and the introduction of computer-based items broadened the construct beyond what was measured in 2009, the last time that reading was a major domain. The framework and instruments will remain constant for the 2018, 2021 and 2024 cycles, with the next revision of the assessment expected for 2027 when reading will again be the major domain. This approach results in more accurate information about item functioning across cycles as the item parameter estimation occurs when a construct is treated as a major domain.

Multistage Adaptive testing

The PGB's long-term development strategy for PISA includes the objective of continuing to exploit the advantages of computer-based testing, including the increased use of adaptive testing to further improve measurement accuracy and efficiency, especially at the extremes of the proficiency scale. Additionally, by allowing measurement across a broader range of the ability distribution, adaptive testing could be viewed as making it possible to include a more diverse set of participants, thereby extending the global reach of PISA.

In 2018, multistage adaptive testing was adopted for the reading literacy domain only. To prepare for multistage adaptive testing, during the PISA 2018 field trial design for reading literacy, unit order was varied to examine whether the order in which units are presented has any impact on item parameter and proficiency estimation. The results of this study in the field trial showed that unit order did not have a significant impact on item parameters nor on proficiency estimates, supporting the use of a multistage adaptive testing design for reading in the main survey. More information about this aspect is provided under the main survey design section of this chapter.

Goals and domain coverage

The design for the PISA 2018 core assessment was developed to provide participants with the following information:

- population distributions in reading that reflect the new 2018 framework as well as links to the framework and proficiency scale developed in 2009
- population distributions in science linked to the 2015 framework and science proficiency scale
- population distributions in mathematics linked to the 2012 framework and mathematics proficiency scale
- population distributions in global competence, the innovative domain in 2018
- pairwise covariance estimates among each of the four domains (reading, science, mathematics, global competence)
- three-way covariance information among the four cognitive domains including the three core PISA domains (reading, mathematics, and science)
- data to link the two modes of delivery: paper-and-pencil and computer-based[1].

In addition to the four core domains of science, mathematics, reading and global competence, the PISA assessment included an optional assessment of financial literacy.

To meet these goals, the items for each domain for the PISA 2018 field trial and main survey were organized as shown in Table 2.2 and Table 2.3. All new items for reading literacy were developed as computer-based items. The reading Field Test design included six clusters of trend items and twelve clusters of new items to study the unit order effects, as a pre-requisite to the introduction of the multistage adaptive testing design in the main survey. Then, in the main survey, the reading items were assigned according to the multistage adaptive design described later in this chapter.

[1] The mode of assessment for most of the participants was computer-based (71 CBA participants), with 9 participants still implementing the PISA 2018 cycle as a paper-based survey.

As shown in Table 2.2, there was no new item development for science and mathematics in 2018. As was the case in the 2015 cycle, mathematics was assessed using existing trend items that were delivered both in paper-and-pencil and computer modes. Because new items for science had been developed for the 2015 cycle, in 2018, the science trend instruments were re-assembled (i.e., different positions in 2015 and 2018) and differed by mode. For countries using the computer-based instruments (see Table 2.2), the trend science items included both 76 items that had been newly developed for the 2015 cycle and 39 items from prior cycles. Finally, global competence items were designed for a limited paper-based pilot study and administration only on the computer for the main survey. Countries using the paper-based instruments (see

Table 2.3) administered only existing trend items taken from cycles prior to 2015 – that is, the 2018 paper-based instruments were identical to those used in 2015.

Table 2.2: Domain coverage for PISA 2018: CBA

Note that each cluster was designed to take approximately 30 minutes of testing time
* Within the six clusters of trend materials, there were two versions of clusters R6 and M6 (6A = standard items; 6B = easier items) but each country used only one version, resulting in six clusters of administered items
** The number of mathematics items in CBA was one less than in PBA because one trend items required students to draw a graph and could not be replicated in CBA
*** Field Trial trend clusters R6A and R6B were not included in the MS adaptive design for reading because such country-level adaptation is not needed for the adaptive design
‡ Note that the global competence items were not part of the 2018 FT but were piloted on paper as part of a separate study in seven countries

Table 2.3: Domain coverage for PISA 2018: PBA (included trend items only)

*There were two versions of clusters R6 and M6 in PBA (6A = standard items; 6B = easier items) but each country/economy used only one version, resulting in six clusters of administered items

OVERVIEW OF THE FIELD TRIAL ASSESSMENT DESIGN

The PISA 2018 field trial was designed to provide the information needed in preparation for the main survey. In particular, it was designed to verify the feasibility of the multistage adaptive test (MSAT) design planned for the main survey. To ensure an appropriate sampling of content and scaling of items and to better adapt to students' performance, the PISA MSAT design offers many alternative options for the selection and delivery of three item testlets among many pre-assembled testlets. As part of the design units need to be assigned to more than one testlet in different positions. Thus, while the order of items within a unit does not change, the position of a unit across testlets can be different. For example, a certain unit can be presented as the first unit in some testlets, but as the second unit in others. Therefore, it is important to verify that the psychometric properties of the items and units are invariant, regardless of their position across different testlets (i.e., absence of item/unit position effects).

The observation of order effects in early PISA cycles has led to the assumption that intact cluster positions are needed for parameter invariance to hold. This is an observation that was unique to PISA and was not found in other large-scale assessments, including PIAAC and the

National Assessment of Educational Progress (NAEP). There was no need to adjust item parameters based on relative item positions in the cognitive instruments. However, a rescaling study conducted on the joint database of all historical PISA data collected between 2000 and 2012 showed good stability of item parameters overall across multiple survey cycles even though over time there were deviations from the strict application of the "intact cluster" paradigm. The 2018 field trial was designed to provide additional information in regard to item parameter invariance under variable unit positions. To that end, the field trial collected data to study unit order effects by manipulating fixed and variable positions within 30-minute (intact) clusters, and students were randomly assigned to three groups with different unit orders.

For the 2018 PISA field trial, a unit was considered to represent the minimum granular size of item sets at which adaptiveness can take place. Units consist of a set of items based on a common stimulus or stimuli that can be considered as the organizing grain size that can be assigned randomly or guided by adaptiveness. While within-unit adaptiveness would be possible in principle, altering the sequence of items within a unit could change the substantive meaning or context, so no variations were introduced within a unit. However, the sequence of units within a cluster can be changed to examine parameter invariance relative to unit position. Examining and ensuring parameter invariance at the unit level was a necessary condition for PISA 2018 to move to adaptive testing.

The goals of the field trial design included:

- evaluation of the invariance of item parameters compared to previous PISA cycles for the 2018 cycle (both CBA and PBA)
- evaluation of the invariance of item parameters regarding the positions of intact item sets (i.e., units); that is, a comparison of stability of item parameters between 30-minute clusters found in prior PISA rounds versus varying positions of smaller collections of units to examine the feasibility of introducing multistage adaptive testing in the main survey
- obtaining preliminary item parameters for the evaluation of new reading and global competence items, and for the selection of a final set of items used in the main survey for these new units
- evaluating sampling and survey operations
- assessing how well the computer platform functions within and across participating countries.

Because the primary goal of the field trial was to support the goals noted above and not to estimate the proficiency distribution of national populations nor item parameters for new items, the sampling requirements differed from those for the main survey.

Like the main survey design, the field trial design for PISA 2018 implemented one CBA design including reading, mathematics, science, and global competence as core domains and financial literacy as the optional domain, as applicable. In addition, the field trial design also included one PBA design that involved the core domains of reading, mathematics, and science as they were implemented in 2015.

The standard design for countries choosing computer delivery was to select a minimum of 28 schools for the field trial and select 64 or 65 students within each school. This resulted in a sample size of 1,800 assessed students. Alternative designs were available for countries having difficulty in finding large schools to implement this design.

Countries that chose to measure student performance with only paper-and-pencil forms had a much-reduced sample size because they used forms from previous cycles. The goals for these countries was mainly focused on testing operations and data-related procedures. These countries selected 25 schools but only selected 36 students from each school for a total field trial sample of 900 assessed students.

Field trial computer-based design

The field trial design needed to support the key goals previously mentioned, including the invariance of item parameters across PISA cycles and modes, and the invariance of item characteristics about the positions of intact item sets (i.e., units). In addition, preliminary item parameters needed to be estimated for the new reading and global competence.

The computer-based assessment (CBA) design organized students into three groups. Students in Groups 1 and 3 took fixed-unit order (FUO) forms, while students in Group 2 took variable-unit order (VUO) forms as shown in Figure 2.1. Each test form consisted of four 30-minute clusters assembled from at most two domains, resulting in at least one hour of assessment time per domain, with a total of two hours of testing time per student. Each cluster consisted of multiple units (item sets), and the ordering of the units were always fixed and consistent in FUO forms in Groups 1 and 3. In contrast, ordering of the units were varied across forms in VUO forms in Group 2. For example, R1 cluster in Form 19 had a different ordering of units compared to the ordering of units in R1 cluster in Form 25. More specifically, Group 1 included Forms 1–18 and measured trend mathematics, reading, and science items. All forms in this group were fixed-unit order forms. Group 2 included 24 forms (Forms 19–42) and measured both new and trend reading items and these forms were variable-unit order forms. Group 3 included 12 forms (Forms 43–54) and measured only new reading items and were fixed-unit order forms. Each student received one of the forms from his or her assigned group.

The field trial standard design is shown in Figure 2.1 where students were randomly assigned to one of these three groups.

Figure 2.1 Field trial computer-based assessment design

Where:

- R1-R6ab represent CBA trend reading clusters
- R7-R18 represent CBA new reading clusters
- M1-M6ab represent CBA trend mathematics clusters
- S1-S6 represent CBA trend science clusters (assembled from 2015 trend and new items)

Field trial paper-based design

Countries that chose the PBA design sampled 25 schools and selected only 36 students from each school for a total field trial sample size of 900. These students were randomly assigned one of the 18 paper-and-pencil forms that contained the trend items from two of the three core domains for PISA – reading, mathematics, and science. This design is shown in Figure 2 2.

Figure 2 2 Field trial paper-based assessment design

Where:

- PR1-PR6ab represent PBA trend reading clusters
- PM1-PM6ab represent PBA trend mathematics clusters
- PS1-PS6 represent PBA trend science clusters (same clusters from 2015)

OVERVIEW OF THE MAIN SURVEY ASSESSMENT DESIGN

The assessment design for PISA 2018 was planned so that the total testing time for measuring the four core domains of reading, mathematics, science, and global competence was two hours for each student. An overview of the flow of the integrated design for the PISA 2018 main survey is presented in Figure 2.3.

Figure 2.3 Overview of the PISA 2018 main survey integrated design

Paper-based integrated design

For PBA countries, the main survey assessment design included 30 forms. These are shown in Figure 2.4. All the items included in the PBA test forms were taken from previous cycles of PISA. Each form included one hour of reading items and items from at least one of the other two core domains. As a result, all students were administered two clusters of reading items, 46% of participating students were administered two clusters of mathematics items, 46% were administered two clusters of science items, and 8% were administered one cluster of mathematics and one cluster of science items, thus providing the covariance information about the three domains. The PBA was to be administered to 35 students in each of 150 schools.

Figure 2.4 Main survey paper-based assessment design

Where:

- *PR01-PR06* represents reading clusters (trend)
- *PM01-PM06* represent mathematics clusters (trend)
- *PS01-PS06* represent science clusters (trend)
- Letters *a* and b following the cluster name represent standard clusters and easier clusters, respectively. These were used only in math and reading.

Computer-based integrated design

For CBA countries including the global competence assessment, the main survey included 72 forms (forms 01-72) that are shown in Figure 2.5. Under the full integrated design, all sampled students responded to 60 minutes of reading items, 41% responded to mathematics items, 41% responded to science items, and 30% responded to global competence items.

For countries not participating in the global competence assessment, only 36 forms were included in the design (forms 01-36) and the percentages for this alternative design are also represented in the first part of Figure 2.5.

Figure 2.5 Main survey computer-based assessment design [Part 1/2]

Figure 2.5 Main survey computer-based assessment design [Part 2/2]

Where:

- *R(adaptive)* represents the reading assessment with the multistage adaptive testing design (trend and new items)
- M01-M06 represent mathematics clusters (trend)
- S01-S06 represents science clusters (trend)
- GC01-GC04 represent global competence clusters (new)
- Letters *a* and b following the cluster name represent standard clusters and easier clusters, respectively. These were used only in math and reading.

Figure 2.6 Overview of the main survey computer-based assessment design

Main survey multistage adaptive design: Reading

The multistage adaptive design (MSAT) that was implemented for reading in PISA 2018 main survey was a modified version of the one which was successfully used in the OECD Programme for International Assessment of Adult Competencies (PIAAC). This section describes the features of the PISA adaptive design.

The MSAT design for the PISA 2018 main survey consisted of three stages and 245 items, each belonging to one of 45 distinct units: 5 units in the Core stage, 24 units in Stage 1 and 16 units in Stage 2. These units were organized into testlets that represent a combination of several of these units with each unit appearing in more than one testlet. Table 2 4 to Table 2 6 show how the testlets in each of the three stages were constructed in terms of number of units, number of items and how units were linked across testlets. The columns represent the "units" and number of total and auto-scored items. The rows represent a the testlets that are comprised of multiple units and number of total and auto-scored items for each testlet. As shown in Table 2 4, the Core stage include 8 different testlets (included a set of five different units, each composed of three to five items. Table 2 5 shows that at Stage 1, there was a set of 24 different units, each composed of three to six items that varied in difficulty from easier ("low" testlet) to somewhat more difficult ("high" testlet). Finally, Table 2 6 shows that at Stage 2, there was a set of 16 different units, each composed of five to eight items across various difficulty levels as well. Testlets at each stage were deliberately linked through common sets of units to ensure the accurate estimation of the item parameters and the proficiency distributions. As a result of this design, each student took 7 units and between 33 and 40 items, depending on which testlet was taken at each stage.

Table 2 4 Main survey computer-based MSAT design of core reading testlets

Table 2 5 Main survey computer-based MSAT design – Stage 1 reading testlets

Table 2 6 Main survey computer-based MSAT design – Stage 2 reading testlets

In the context of this MSAT design, there were parallel testlets at each of the three stages: Core (eight testlets labelled RC1-RC8 that were split into two parallel sets), Stage 1 (eight more difficult testlets labelled as R11H-R18H for "high" and eight easier testlets labelled as R11L-R18L for "low") and Stage 2 (eight more difficult testlets labelled as R21H-R28L for "high" and eight easier testlets labelled as R21L-R28L for "low"). These parallel sets are illustrated in Figure 2 7 by the two shaded green areas for the Core and two yellow shaded areas for Stage 1. The introduction of parallel sets: i) doubled the number of items to ensure better coverage of the domain and thus, representation of the scale, and ii) served as the linking unit as testlets are also linked across "parallel sets" at every stage. More specifically, shaded cells in Table 2 5 demonstrate how parallel sets were linked to each other: the patterns linking the sets from

R11H to R14L to Stage 2 were almost symmetric with the patterns linking the sets from R15H to R18L to Stage 2.

Figure 2 7 Routing paths in the standard computer-based MSAT design – Design A – that connect Core>Stage 1>Stage 2 (64 paths in total, applicable to 75% of students)

With respect to routing, at the Core stage, testlet assignment was based on a random number between 1 and 8. At Stage 1, testlet assignment was based on three criteria: i) the Core testlet assigned, ii) the students' performance on the Core (i.e., total number correct on auto-scored items on the given testlet), and iii) a probability layer matrix. Similarly, at Stage 2, testlet assignment was based on: i) the testlet taken at Stage 1, ii) the performance at Core and Stage 1 (i.e., total number correct on auto-scored items on the given Core testlet and Stage 1 testlet), and iii) a probability layer matrix.

More specifically, the routing method depended on performance on the auto-scored items at the previous stage, which categorized students into three performance groups: low, medium and high. Decisions about these performance groups were based on the same proficiency thresholds corresponding to testlet-specific number correct on the auto-scored items. For each testlet, the number-correct thresholds were computed using their test characteristic curve (TCC). The number correct score corresponding to the PISA scale scores of 425 and 530 were used to set the lower and the upper thresholds, respectively. This method using the testlet-dependent total number correct thresholds was used to optimize the expected gains from the multistage adaptive testing procedure, considering that at each decision point, only partial information was available. Performance on human-coded (HC) items were not used in drawing the TCC, and therefore were not considered in the routing decision.

Appendix 1 of this chapter provides an overview of the adaptive process for the standard design, or Design A, that was applicable to 75% of the respondents (shown in Figure 2.7). The first part of the table (green columns) shows the Core testlets (RC1 to RC8) along with the number of auto-scored items and the number-correct thresholds, and how the process was applied to select Stage 1 testlets.

- For example, a student taking Core 1 was classified as the "low" performance group if the number-correct was lower than 4, as the "medium" performance group if the number-correct was between four and six, and as the "high" performance group if the number-correct was more than six. It also shows that a student in the "low" performance group had a 90% probability of being assigned an easy testlet (labelled as "L") in Stage 1 (in the case of Core 1, the assigned testlet would be either R11L or R15L) and a 10% probability of being assigned a difficult testlet (labelled as "H") in Stage 1 (again, in the case of Core 1, the assigned testlet would be either R11H or R15H). The opposite happened for students in the "high" performance group at Core. However, a student classified as "medium" performance group at Core had an equal probably of being assigned any of the two difficult or two easy testlets in Stage 1.

The second part of the table (yellow columns) describes the Stage 1 testlets in combination with the Core testlets. In this context, the total number of auto-scored items was a sum of the Core + Stage 1 testlet which was the most complete level of information available at this point for each student.

- For example, a student who took Core 1 testlet and was assigned testlet R11H in Stage 1, was classified as the "low" performance group if the number-correct was less than eight, as the "medium" performance group if the number-correct was between eight and 13, and as the "high" performance group if the number-correct was more than 13. It also shows that a student in "low" performance group (combined with Core and Stage 1) had a 90% probability of being assigned an easy testlet in Stage 2 (in this specific example, the assigned testlet would be R21L) and a 10% probability of being assigned a difficult testlet in Stage 2 (R21H). The opposite happened for students in the "high" performance group (combined with Core and Stage 1). However, a student in the "medium" performance group (combined with Core and Stage 1) had an equal probably of being assigned a difficult or an easy testlet in Stage 2.

Finally, the third part of the table in Appendix 1 (white columns) describes the Stage 2 testlets and their association with the combination of Core and Stage 1 testlet.

To control for the possible position effects of items located at Stage 2 and to increase the accuracy of parameter estimation for all items, an alternative design was added to provide additional routing paths, Core>Stage 2>Stage 1 or Design B, that was applicable to 25% of the respondents (shown in **Error! Reference source not found.**). These alternative routing paths added connections between difficult and easy parallel testlets between Core and Stage 1, which doubled the number of paths from 64 in the standard design to 128 paths in the alternative design. Note that the 128 paths in the alternative design still involve the 64 paths in the standard design. More specifically, in the standard design (Figure 2 7), a Stage 2 difficult testlet was only routed from Stage 1 difficult testlets (e.g., R21H was routed from R11H and R15H). In the alternative design (Figure 2.8), Stage 2 difficult testlets were routed to not only Stage 1 difficult testlets but also to easy testlets (e.g., R21H is routed to R11H, R15H, R11L, and R15L). Therefore, the final design was the combination of Design A (for 75% of students) and Design B (for 25% of students).

Figure 2.8 Routing paths in the alternative computer-based MSAT design – Design B – that connect Core>Stage 2>Stage 1 (128 paths in total, applicable to 25% of students)

Une heure (UH) form

Consistent with previous cycles, a special one-hour test, referred to as the "Une Heure" (UH) form, was prepared for students with special needs. The selected items were among the easier items in each domain and had a more limited reading load. The UH form contained about half as many items as the other instruments, with each cluster including from seven to nine items. The UH form was comprised of about 50% reading, 25% mathematics and 25% science items.

The UH form included two 15-min clusters of reading (RU1 and RU2), one 15-min cluster of mathematics (MU1) and one 15-min cluster of science (SU1). The assignment of this booklet followed the approach described previously for the assignment of the base test form. The UH form was assigned base form 99 (as shown in Figure 2.9).

Figure 2.9 Main survey UH form design

The UH form was accompanied by a UH student background questionnaire that included a subset of items from the regular background questionnaire (primarily trend items) in a single form design that was administered in CBA only, as no PBA participants chose to administer the UH Form.

Assessment of financial literacy

The assessment of financial literacy was offered as an international option in PISA 2018. In total, 21 countries opted to administer these instruments. The cognitive instruments included trend items from 2012 and 2015 plus a set of new interactive items that were developed specifically for PISA 2018. Financial literacy was available only as a computer-based assessment because participants in this option were all implementing PISA as a computer-based survey.

Financial literacy was administered to a separate sample of PISA-eligible students who took a combination of reading, mathematics, and financial literacy items during a total testing time of two hours (120 minutes) for each student. The group of students who took financial literacy are referred to as "Financial Literacy sample". Note that this was different from the approach used in the 2015 cycle, when FL was administered to a subset of the students in the main sample.

Field trial financial literacy design

For the field trial, the main sample was augmented by adding a sample of approximately 384 students who were assigned one of the 12 financial literacy testing forms. These forms included 60 minutes of financial literacy items, 30 minutes of trend reading items and 30 minutes of trend mathematics items. They were based on three clusters of financial literacy (FL1 to FL3) and six trend clusters each reading (R1 to R6ab) and mathematics (M1 to M6ab). The design is shown in Figure 2.10. The 12 financial literacy forms were administered to Group 1 (FUO) and each form was taken by about 32 students within each country.

Figure 2.10 Field trial computer-based financial literacy design

Main survey financial literacy design

For the main survey, countries administering the financial literacy instruments required 1,650 additional students. Each student taking the financial literacy assessment took financial literacy items, and mathematics or reading items, with the reading items administered in the same adaptive mode as in the main sample, including the fluency tasks. Students taking the financial literacy assessment did not take any of the science items and therefore they do not have science literacy proficiency estimates.

The main survey version of the assessment instruments included 43 financial literacy items, of which 29 were trend items and 14 were new items. These items were organized into two 30-min clusters of financial literacy (FL1-FL2) that were rotated into twelve forms each containing 60 minutes of financial literacy and 60 minutes of mathematics or reading as shown in Figure 2 *11*.

Figure 2 11 *Main survey* computer-based financial literacy design

THE 2018 READING ASSESSMENT 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 target distribution of items according to important framework variables – and outline the possibilities for reporting results. For PISA 2018, subject matter expert groups (SMEGs) were convened by the Core B contractor to develop a framework for reading literacy.

Over the course of the project, two different expert groups worked on global competence — the first convened by Core B and the second by the OECD, following a change in the strategy for the global competence domain. The scientific and mathematical literacy frameworks were based on those developed for the 2015 and 2012 assessment cycles, respectively, when these domains were major domains.

Reading Literacy Cognitive Processes

The reading literacy domain describes reading in terms of cognitive processes.

Successful reading, whether reading a single text or reading and integrating information across multiple texts, requires an individual to perform a range of processes. The 2018 reading literacy framework defines several cognitive processes that span a range of difficulty. Each cognitive process is assigned to a superordinate category which will be used for the final scaling of the 2018 main survey data: Locate information, Understand, and Evaluate and Reflect. The cognitive processes within each category are briefly defined below.

Locate information

- Access and retrieve information within a text – scanning a single text in order to retrieve target information consisting of a few words, phrases or numerical values.
- Search for and select relevant text – searching for information among several texts to select the most relevant text given the demands of the item/task.

Understand

- Represent literal information – comprehending the literal meaning of sentences or short passages, typically matching a direct or close paraphrasing of information in the question with information in a passage.
- Integrate and generate inferences – going beyond the literal meaning of information in a text by integrating information across sentences or even an entire passage. Tasks that require the student to create a main idea or to produce a summary or a title for a passage are classified as “integrate and generate inference” items.
- Integrate and generate inferences across multiple sources – integrating pieces of information that are located within two or more texts.

Evaluate and Reflect

- Assess quality and credibility – evaluating whether the information in a text is valid, current, accurate, unbiased, reliable, etc. Readers must identify and consider the source of the information and consider the content and form of the text or in other words, how the author is presenting the information.
- Reflect on content and form – evaluating the form of the writing to determine how the author is expressing their purpose and/or point of view. These items often require the student to reflect on their own experience and knowledge to compare, contrast or hypothesize different perspectives or viewpoints.
- Detect and handle conflict – determining whether multiple texts corroborate or contradict each other and when they conflict, deciding how to handle that conflict. For example, items classified as "detect and handle conflict" may ask students to identify whether two authors agree on the stance of an issue or to identify each author's stance. In other cases, these items may require students to consider the credibility of the sources and demonstrate that they accept the claims from the more reliable source over the claims from the less reliable source.

Reading Literacy Texts

Reading texts can be classified along four different dimensions described in the framework: source, organization and navigation, format, and type. Each dimension is briefly described below.

Source

- Single – a single unit of text that has an author or a group of authors, a time of writing or publication date and a reference title or number.
- Multiple – multiple units of texts where each has a different author, different publication times or have different titles or reference numbers.

Organization and Navigation

- Static – texts with simple organization and a low density of navigation tools; typically texts with one or several pages organized in a linear way.
- Dynamic – texts with a more complex, non-linear organization and a higher density of navigation tools.

Format

- Continuous – texts formed by sentences that are organized into paragraphs.
- Non-continuous – texts composed of a number of lists or elements such as tables, graphs, diagrams, advertisements, schedules, catalogues, indexes, forms, etc.
- Mixed – texts containing both continuous and non-continuous elements.

Type

- Description – texts with information that refers to properties of objects in space. Description texts provide an answer to "what" questions. Examples include a depiction of a place in a travelogue, a catalogue or a process in a technical manual.

- Narration – texts with information that refers to objects in time. Narration texts provide answers to “when” or “in what sequence”. Examples include a report, a news story, a novel, a short story or a play.
- Exposition – texts with explanations of how different elements interrelate in a meaningful way and provide answers to “how” questions. Examples include a scholarly essay, a diagram showing a model of memory, a graph of population trends, or a concept map for an entry in an online encyclopedia.
- Argument – texts that present the relationship among concepts or propositions. Argument texts provide answers to “why” questions. An important subclassification of argumentative texts is persuasive and opinionative texts, referring to opinions and points of view. Examples include a letter to the editor, a poster advertisement, posts in an online forum or a review of a book or film.
- Instruction – a text that provides instructions on what to do. Examples include a recipe, a series of diagrams showing how to give first aid or guidelines for operating software.
- Transaction – a text that aims to achieve a purpose such as requesting that something is done, organizing a meeting or making a social engagement with a friend. Examples include a letter, an email or a text message.

Reading Literacy Scenarios

Reading is a purpose-driven activity; that is, it occurs when a reader wishes to accomplish a particular goal, such as locating information to fill out a form or understanding a topic well enough to participate in a discussion with peers. In many traditional reading assessments, however, the “goal” is simply to answer a few discrete questions about a text on a general topic and then move on to the next. In contrast to this somewhat artificial world of traditional reading assessments, the reading units developed for 2018 are scenario-based. Each unit begins with a fictional scenario that describes the over-arching goal for reading the text or collection of texts in the unit. Thus, the reader is given both a context and a purpose that helps to shape the way he or she searches for, comprehends, and integrates information.

Scenarios were developed to address a range of situations. The framework describes several types of situations that relate to the overarching scenario developed for each unit:

- Personal – situations that contain text that satisfies an individual’s personal interests in both practical and intellectual ways. Examples include personal letters, fiction, biography and informational texts that are read to satisfy curiosity or for leisure as well as personal emails, instant messages and blogs.
- Public – situations that contain text that relates to activities and concerns of the society at large. Examples include official documents, information about public events, message boards, news websites and public notices.
- Educational – situations that contain text designed for the purpose of instruction and that is often chosen by an instructor rather than the reader. Examples include printed or electronic textbooks and interactive learning software.
- Occupational – situations that contain text that supports the accomplishment of an immediate task. Examples include texts used to search for a job such as printed classified ads or online job websites, and texts that provide workplace directions.

Reading Fluency

A measure of reading fluency was included in PISA 2018 to provide additional information about students' reading ability towards the lower end of the reading proficiency scale. PISA defines reading fluency as the ease and efficiency with which one can read text. In this task, students were presented with one sentence which either made sense or not. Students were instructed to respond "Yes" if the sentence made sense and "No" if the sentence did not make sense. As soon as the student responded, the next sentence appeared. Students were told they would have three minutes to respond to as many sentences as they could.

The ability to read text fluently is an important skill that supports a reader's ability to comprehend text more easily and competently. When readers can efficiently decode and recognize words, access their meaning and parse syntax, it frees up cognitive resources to focus on higher-level aspects of comprehension. When readers cannot read text fluently, resources must be allocated to lower-level reading processes which limits the readers' ability to engage with text at a higher level.

The goal of adding reading fluency to the assessment of reading literacy in PISA 2018 was to be able to better characterize the reading ability of students at the lower proficiency levels.

THE 2018 GLOBAL COMPETENCE ASSESSMENT FRAMEWORKS

As the innovative domain for the 2018 cycle, the global competence assessment focused on the skills that twenty-first century students need in today's increasingly interconnected, diverse and rapidly changing world. The domain was defined as follows:

> *Global competence is the capacity to examine local, global and intercultural issues, to understand and appreciate the perspectives and world views of others, to engage in open, appropriate and effective interactions with people from different cultures, and to act for collective well-being and sustainable development.*

Four cognitive processes that support global competence were further defined and included:

- The capacity to **evaluate information, formulate arguments, and explain complex situations and problems** by using and connecting evidence, identifying biases and gaps in information and managing conflicting arguments.
- The capacity to **identify and analyse multiple perspectives** and world views, positioning and connecting their own and others' perspectives on the world.
- The capacity **to understand differences in communication**, recognising the importance of socially-appropriate communication conventions and adapting communication to the demands of diverse cultural contexts.
- The capacity to **evaluate actions and consequences** by identifying and comparing different courses of action and weighing these actions against one another on the basis of short- and long-term consequences.

Figure 2.12 illustrates the relationship between the cognitive skills and knowledge to be assessed and the dimensions of global competence, as those are identified in the domain definition.

Figure 2.12 The relationship between the cognitive test of global understanding and the dimensions of global competence

ROLE OF THE SUBJECT MATTER EXPERT GROUPS IN ITEM DEVELOPMENT

As the contractor for instrument development, Core A was responsible for working with the subject matter expert groups (SMEG) as applicable.

For reading, Core A worked with the reading Expert Group (REG) to understand their vision for the range and types of items to be developed for PISA 2018. To facilitate the transition from the work of Core B (framework development) to the instrument development activities, Core A retained the SMEG members who met under Core B to begin work on the frameworks in September 2014 and January and June 2015. For global competence, the OECD developed the final framework and worked with Core A under a revised strategy whereby the first meeting took place in June and September 2017 to finalize the main survey Item Pool.

Core A's work with the SMEGs began in June 2015 and focused on the following tasks:

- describing the kinds of items needed to assess the skills and abilities in each domain as defined in the framework
- reviewing and understanding the proposed assessment design in order to define the number and types of items that were needed for each of the domains
- defining the behaviours of interest for the computer-based tasks
- defining the intersection between the kinds of functionality that might be desirable for measuring the constructs and the functionality that was practical to implement in the assessment.

Work with the subject matter experts continued beyond the initial meetings through instrument development and data analysis. For reading and global competence, SMEG members played an important role in reviewing assessment tasks as they were developed, providing input into the analysis of the field trial data, approving the set of items for the main survey, and working with development and analysis staff to develop the described scales used for reporting the PISA 2018 results.

PISA 2018 TEST DEVELOPMENT

Test development for the PISA 2018 cycle began in early-2015 and focused on the development of items for a computer-based assessment.

Computer-based assessment: Screen design and interface

Multi-page stimulus materials

The screen design and interface developed for the 2015 cycle was used for the 2018 cycle. In 2015, paper-based trend units with long stimulus materials were adapted to computer-based formats using a paging interface that allowed students to move from page to page throughout the text. The paging interface maintained the same kinds of interactions students would have encountered with long texts in the paper-based format. The 2018 reading literacy framework focused on expanding the construct and the new item pool to include more representations of reading in digital environments. This included both an increase in the representations of digital reading formats as well as the tasks and cognitive processes readers encounter when reading digital texts. To support the framework, the design for new reading units used scrolling for viewing longer texts instead of paging and used tabs to present multiple sources within a unit. For units with multiple sources, students could switch back and forth between texts using the tabs at the top of the stimulus pane in a way very similar to how readers use tabs within an internet browser to view multiple pages. The Reading Expert Group supported the switch to scrolling and tabs for new reading units and agreed that it was important to represent digital forms of reading in authentic formats. An example of how this was implemented in PISA is included in Figure 2.13 from the released unit Rapa Nui.

Figure 2.13 Tabs and scrolling display

Several safeguards were included to ensure that students saw all the pages in each unit and understood how to navigate between them. First, students were introduced to the tab and scrolling interface in the orientation. In most multiple source units, the sources were presented one at a time. For example, the first item would appear alongside the first source, similar to how items are presented for single source units. Prior to introducing the second or third source, students were given an update to the scenario to introduce the purpose for next source. In some units, students could access a second or third source by clicking on a hyperlink in the first source. However, if a student did not access the next source by clicking on the link, when the first item that required the next source was presented, the new source would appear in the tab. In this way, all students were given an opportunity to read the source required to complete the item.

Two units had a simulated website with multiple pages, accessible via menu options that appeared at the top of the website's pages, functioning in the same way that tabbed menus work on real-world websites. This format was reviewed in the orientation for reading so that students would be familiar with this design.

Navigation

As in PISA 2015, students could navigate through the items as needed. For most units, students were able to move back and forth between items *within* a unit. They were not, however, able to move back and forth *between* units. Once students clicked on the "NEXT" button on the final item in a unit, a dialog box displayed a warning that the student was about to move on to the next unit and it would not be possible to return to previous items. At this point, students could either confirm that they wanted to go on or cancel the action and continue with the unit on which they had been working.

Response modes

Across all domains, PISA 2018 included items requiring one of five different response modes:

- click on a choice: single-selection multiple choice; multiple-selection multiple choice (click on one or more responses); complex multiple choice (table with statements and a number of yes/no or true/false options); or click on an image
- numeric entry (only numbers, commas, periods, dashes and backslashes could be entered)
- text entry (within a scrolling text box that did not constrain the length of a student response – consistent with what was possible for paper-and-pencil items)

- select from a drop-down menu
- drag and drop (including use of a slider).

Orientations

A general orientation introduced students to the screen design and those response modes that were common across most domains. Students received this orientation before beginning the test. Prior to beginning each section of the test, students received a very short domain-specific orientation with instructions specific to the domain in that section. For example, before beginning the reading section of the assessment, students were introduced to the paging, scrolling and tab interface for the longer stimulus materials.

Trend items

The assessment design for PISA 2018 required that six 30-minute clusters of trend items be taken from previous cycles for reading, mathematics and science. These were the same trend clusters that were used in the 2015 main survey and remain intact between the 2015 main survey and the 2018 field trial. The trend clusters were also the same across paper- and computer-based modes. For the main survey, the reading trend item pool included 72 items, from which 36 required human coding. Items from clusters 6A (standard units) and 6B (easier units) were not included in the main survey because countries that typically administer cluster 6A do not have translations for cluster 6B and vice-versa. Thus, for the adaptive reading design in the main survey, only those units for which all countries had translations could be used.

The mathematics trend item pool for CBA included 82 items and for PBA included 83 items, from which 22 items require human coding. The assessment of financial literacy included 43 trend items from the 2015 main survey, from which 13 were human coded.

The assessment of paper-based science included the same six-clusters used in PISA 2015 with 85 items. For computer-based science, the six trend clusters were re-assembled from the 2015 main survey item pool and included 115 items from which 39 were developed prior to 2015 (that were also in the PBA assessment in 2018) and 76 items were newly developed for the PISA 2015 survey, from which 32 were human coded.

New items

In PISA 2018, test development occurred for the domains of reading, global competence and financial literacy. To prepare for the implementation of the multistage adaptive design in the main survey, twelve 30-minute clusters of new items were developed for reading during the field trial. In total, 31 new units with 254 reading items were selected and included in the field trial. Sixty-five reading fluency items were developed by test developers and assigned to six blocks. Each student received two blocks of sentences which were fully rotated. In the field trial, two blocks of reading fluency items were always administered before two clusters of reading literacy items. For financial literacy, twenty new items were developed for the field trial in cooperation with the Directorate for Financial and Enterprise Affairs within the OECD.

The development and implementation of global competence followed a revised strategy. Rather than conducting a computer-based field trial, global competence was trialled through a paper-based pilot study with five 30-minute clusters. The global competence item pool included 19 units with 86 items in the field trial, from which 21 items were human scored. These items were developed in cooperation with the OECD.

International test development team

Test development efforts were coordinated by ETS as the Core A Contractor. As is the case with any large-scale international survey, it is important that the pool of tasks used in PISA reflect the range of contexts and experiences of students across participants. One way to meet this goal is by convening an international team of item developers. For PISA 2018, the international test development team included individuals from the University of Luxembourg and the University of Liège. A second way to meet this goal was to work with countries or economies on submission and development of materials. Core A provided countries with a range of opportunities for participation during the development process.

National submissions

The active involvement of countries or economies in the development process is important for the data-collection instruments to be internationally valid and representative. Thus, it is important to ensure that the final item pool reflects the international context of an assessment such as PISA. PISA 2018 included a range of approaches and multiple opportunities for countries or economies to participate in the development process. Given the nature of reading, Core A decided to split this process of national submission into two phases:

- A stimulus submission phase that took place between February and August 2015. During this phase, 14 countries/economies submitted stimuli and 16 of the 31 field trial units either used stimuli submitted by countries/economies or were inspired by their submissions.
- An item submission phase that took place between April and October 2015. During this process, countries/economies expressed an interest to participate in this process and received stimuli from Core A to develop items. A total of 15 countries/economies contributed items during this phase.

Global Competence

The test development process for global competence followed a different plan. For the revised strategy of this domain, the OECD worked with participating countries to create scenarios to which items would be written. Through a process of country submission and country review, the OECD selected a set of 20 scenarios. Core A test developers further developed the scenarios by simplifying the text when necessary and in some cases, adding stimuli to the scenario to create more opportunities for items. The OECD reviewed all scenarios and items early in the review process, prior to country reviews to ensure the items fulfilled the goals of the revised framework. Once approved, the units received the same reviews conducted for new reading units.

Item development workshops

Three item development workshops were offered as part of the PISA 2018 efforts to involve participants in the test development process. These took place in April and May 2015 in Washington D.C., Barcelona and Singapore. From the test developers' view, the workshops made the development process more efficient because of the in-person training and collaboration, which was reflected in the quality of items that came out of the workshop and the items that were submitted subsequently. These workshops allowed national representatives to interact and share ideas and expertise with members of the test development teams. Countries wrote and reviewed each other's items at the workshop, and received some "real-time" feedback from the test development teams. The workshops also provided a venue to exchange ideas for ways to assess newly developed content at the lower end of the student performance scale.

Fifty-five participants from 28 countries/economies attended these workshops. They were extremely successful and produced 157 items for eight units. Countries could, of course, still submit items outside these workshops and, if favourably reviewed, they were incorporated into the item pool development.

Item Reviews

Newly developed units were submitted for translatability review at the same time they were released for country review. Linguists representing different language groups provided feedback on potential translation, adaptation and cultural issues arising from the initial wording of items. Experts at cApStAn and the translation referee for the 2018 cycle alerted item developers to both general wording patterns and specific item wording that are known to be problematic for some translations and suggested alternatives. This allowed item developers to make wording revisions at an early stage, in some cases simply using the alternatives provided and in others working with cApStAn to explore other possibilities.

All reading and financial literacy items were released for country review prior to the field trial. Countries had two weeks to preform reviews and submit feedback on all draft stimuli and items. Reading items were released in three batches. Test developers received review forms from 42 countries for Batch 1, 49 countries for Batch 2 and 49 countries for Batch 3. The financial literacy items were released in one batch which was reviewed by 13 countries.

Global competence items were released for country review on a different timeline than the reading and financial literacy items prior to the field trial pilot. However, 23 countries reviewed the draft stimuli and items prior to the finalization of the global competence instruments.

Preparation of the French source version for all new units provided another opportunity to identify issues with the English source version related to content and expression. Development of the two source versions helped ensure that items were as culturally neutral as possible, identified instances where wording could be modified to simplify translation into other languages, and specified where translation notes would be needed to ensure the required accuracy in translating items to other languages.

In addition, user testing was conducted by the University of Liège focusing on navigational challenges and user interface. These focused on the navigation features used across units with multiple texts, the tab design for multiple source units, and paging (the approach used in 2015) vs. scrolling. Students found tabs and scrolling to be more natural and more like real world digital experiences. Cognitive labs were implemented by the University of Luxembourg focusing on the digital reading designs and timing. These sessions observed interactions with digital elements and timing, how well students interacted with the designs, revealed digital design elements that were then highlighted in the orientation, and provided input into the number of items to include in the multiple text units for the field trial.

Information from these sessions was used to revise one interface element in reading and correct several identified bugs. Equally important, the questions raised by study participants informed the development of the domain orientations, identifying areas where students needed instruction and practice before working on the assessment items.

Selection of new items for the field trial

The 2018 item development process produced a total of 34 reading units with 318 items. Items were selected for inclusion in the field trial based on country reviews, feedback from the expert group and the distribution of items across the key categories as defined in the framework. Of these 318 reading items, 254 items from 31 units were selected by the Reading Expert Group for the field trial. Of these 254 items, 53 percent originated from the national submissions received from 17 countries or economies. An additional 23 percent of the field trial items were originally developed at the Item Development Workshops, and 24 percent of the items were developed solely by ETS's test development team.

Nineteen global competence units with a total of 86 items were developed for the paper-based pilot test that took place in seven countries as part of the alternative strategy implemented by the OECD. Items were selected for the FT based on country reviews and unit coherence, construct coverage. Following the field trial pilot, test developers made a selection of items for the MS, balancing item quality, item information and construct coverage and presented this to the expert group. The expert group reviewed all the items and made suggestions for the inclusion of some items and recommended the inclusion of one new item. They worked with the test development team to construct the item and coding guide. All pilot items were adapted to computer format (following the paper pilot) in preparation for the expert group. Once the selection was finalized, the items not selected were removed from the CBA. Coding guides were revised using the information collected during coder queries for the pilot. Coding guides were further revised following the international coder training in Malta.

FIELD TRIAL

The PISA 2018 field trial data collection timeline began in March 2017 and extended through August 2017 with 80 participating countries or economies across 140 language versions. Across modes, 71 participants implemented PISA as a computer-based assessment while nine participants implemented it as a paper-based assessment. 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 2018, the Core A contractors continued to assume responsibility for assembling the assessment instruments for both paper- and computer-based participants. Participants were responsible for translating all new material and performing both linguistic and layout quality control checks for trend and new items.

Computer-based trend items

Existing computer-based participants, that is, those that implemented PISA 2015 as a computer-based assessment, were given access to the existing XLIFFs from PISA 2015 and were given the opportunity to review their materials for any errors or necessary updates.

For countries switching from paper- to computer-based assessment, the Core A contractors copied those into the computer-readable XLIFF format used for the computer-based instruments. This was done both as a quality control process and to reduce the tasks assigned to countries given the short development timeframe for the project. Once the XLIFF files were created, countries were asked to perform a review by comparing the new computer versions with PDF files of their paper-based items that were supplied by the contractors.

In both cases, countries were asked to document any content errors, which included typographical mistakes or text errors introduced in the process of copying and pasting across formats. Any content issues identified by countries were reviewed by verifiers on the linguistic quality control team and, if approved, the verifiers made the needed change in the computer files. If countries identified any serious layout issues, those were reviewed and, where appropriate, corrected by the Core A technical team. As an additional quality control check, the Core A contractor also performed layout checks of all items in all languages to identify errors that may have been missed.

Computer-based new items

All new reading and financial literacy items needed to be translated by national teams following the translation and reconciliation processes defined in the PISA standards. Following verification of national translations and the corrections of any remaining errors, countries were asked to sign off on their cognitive materials and those files were then considered locked. Countries that participated in the field trial pilot for global competence translated the English sources of the items and coding guides for the paper-based instruments.

Preparing the field trial national student delivery systems (SDS)

The Student Delivery System (or SDS) was a self-contained set of applications for delivery of the PISA 2018 CBA assessments and computer-based student background questionnaires. A master version was assembled first for countries to test within their national IT structure. This allowed countries to become familiar with the operation of the SDS and to check the compatibility of the software with computers being used to administer the assessment.

Once all components of national materials were approved and locked, the national SDS was assembled and tested first by the Core A technical team. The SDS was then released to countries for national testing. Countries were asked to check their SDS following a specific testing plan provided by Core A and to identify any residual content or layout issues. Where issues were identified, those were corrected and a second SDS was released. Once countries signed off on their national SDS, their instruments were released for the field trial.

Paper-based instruments

National versions of the paper-based trend clusters were prepared by the Core A contractor. To better ensure comparability of the paper-based assessment materials across countries and languages, booklets were centrally created by Core A and then reviewed and approved by countries. Those countries who were new to PISA or who were missing some items from previous cycles needed to translate those materials following the standard translation and verification process. All countries needed to update and translate the common booklet parts, which included the cover, general instructions, formula sheet for mathematics and the acknowledgements page.

The approved clusters were then assembled into the 18 field trial paper booklets by the contractors in a centralised approach 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.

Field trial coding

Coding guides for trend items were compiled by Core A based on previous national versions. For items that were in both the computer and paper instruments, both item IDs were included and, where question wording differed between the paper and computer formats, both versions were shown. Any items where the paper version was human coded, but the computer version was automatically scored were also identified.

The development of the coding guide for one batch of new items in reading was informed by cognitive labs conducted by the University of Liège. Coding guides for the remaining new reading items were informed by the information collected in the cognitive labs. The English master version of the new reading coding guide was released in draft form prior to the coder training meeting in January 2017. Based on discussions at that meeting, the coding guide was finalised and the updated English version, along with the French source version, was released to countries in March 2017, prior to the beginning of the field trial data collection period. The coding guide for global competence was developed by test developers at ETS for the field trial pilot. A coder training webinar was held with all participating field trial pilot countries prior to coding. Based on input received in the webinar, the coding guide was finalized in April 2017 in time for the field trial pilot countries to translate the coding guide.

Field trial coder training

The international field trial coder training was held in January 2017 and focused on reading, as the major domain. Coder training for mathematics and science trend items and all financial literacy items was done through recorded trainings. The goals of the training included both having attendees (master coders) 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 an 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 during previous cycles, Core A set up a coder query service for the 2018 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 were provided by domain-specific teams including item developers and, for trend items, by members of the response team from previous cycles.

In addition to responses to new queries, the queries report included the accumulated responses from previous cycles of PISA. This helped foster consistent coding of trend items across cycles. The report was regularly updated and posted for National Centres on the PISA portal as new queries were received and processed.

Field trial outcomes

The PISA 2018 field trial was designed to yield information about the quantity and quality of data collected as well as to prepare the multistage adaptive testing design for the main survey. More specifically, general 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 operational characteristics of the computer-delivery platform
- the quality of the items including both those items that were newly developed for computer-based delivery and those that were adapted from earlier cycles
- the use of the data to establish reliable, valid, and comparable scales based on item-response theory (IRT) models both in paper- and computer-based versions.

Overall, the field trial achieved all the stated goals. This information was crucial for the selection and assembly of the main survey instruments and for refining survey procedures where necessary. Furthermore, the field trial results confirmed the feasibility of introducing multistage adaptive testing in the main survey as unit order effects were found to be negligible.

The field trial analyses were conducted in batches based on data submission dates. Most of the analyses implemented to evaluate the goals noted above were based on data received from countries by 31 July 2017. That included 53 participants, with three from countries implementing only the paper-based assessment and 50 from participants using the computer-based assessment. The initial FT analyses were based on these data representing 35 language groups and 72 country-by-language groups. The field trial analyses were amended after receiving additional data, which increased the number of countries to 80 participants by the end of 2017, including nine participants that implemented PISA as a paper-based survey and 71 that implemented PISA as a computer-based survey.

MAIN SURVEY

The PISA 2018 main survey was conducted between March and December 2018. The majority of countries completed the main survey data collection by May. 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 Core A 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 July 2017 and concluded in October 2017 and was conducted in two phases. The first phase occurred before countries received their field trial data and the second after receipt of their data. This two-phase process was implemented to allow for the most efficient correction of any remaining errors in item content or layout given the extremely short turn around period between the field trial and main survey.

Phase 1 allowed countries to report any linguistic or layout issues that were noted during the field trial, including errors to the coding guides. All requests were reviewed by Core A and assigned to one of two categories: serious errors that would be expected to impact item functioning and therefore were corrected immediately; and comments that would be re-evaluated based on the field trial data. Errors in the first category were corrected centrally by the contractors.

Following release of the field trial data, countries received their Phase 2 updated item feedback forms that included flags for any items that had been identified as not fitting the international trend parameters. Flagged items were reviewed by national teams. As was the case in Phase 1, countries were asked to provide comments about these specific items where they could identify serious errors. Requests for corrections were reviewed by Core A and, where approved, implemented.

Item selection

Reading

The initial selection of items for the main survey was made by the test development team based on item statistics from the first batch of the field trial data. The first step was to evaluate the item functioning of complex multiple-choice items that had a table format. Each row of a table item had a binary response (Yes/No, True/False, etc), so rows that had P+ values close to or less than 50% were eliminated. Rows with a negative r-biserial correlation (a correlation between performance on an individual item with the total score on a cluster of items) or a correlation close to 0 were also eliminated. Scoring rules were developed for complex multiple choices items based on performance across all rows in a table. In some cases, several items were combined into polytomous items with partial credits whereas in other cases, if all rows in a table item were easy, combined items were scored as dichotomous items such as full credit for all correct answers. No full items were eliminated at this stage of the item selection.

Once scoring models were created for the complex multiple-choice items, the items were scaled using IRT. At the second stage of item selection, items with P+ < .20 or > .90, item-skill correlations < .20, IRT slope < .40, number of item-by-country interactions > 10% of countries with misfit, and the number of countries with low coder reliability > 10% could be eliminated for not functioning well in the field test. Nineteen of the 254 field trial items were eliminated based on these criteria.

Finally, at the third stage of item selection, units were shortened such that the units would fit into the multistage adaptive design. Constraints of the design are described in the next section. Shorter units were to have approximately five items while longer units were to have seven. Items could not be selected based on their statistics alone. Instead, items were selected in this final stage such that all the items in the unit maintained good continuity and made good use of the texts within the unit. An effort was made to reduce the number of items that were categorized with cognitive processes that were already represented in the trend units and preferentially select items that represented the new cognitive processes. Then, when all these selection criteria were met, items that had stronger statistics were kept over those with weaker statistics. At this final selection stage, 50 items were eliminated.

Assigning units to the multistage adaptive design

The multistage adaptive design for reading imposed several constraints that had to be balanced with the goals of the framework during item selection for the main survey. These constraints are described here to provide a context for the item selection conducted by test developers.

- Core stage: For test assembly, the goal was to have approximately 10 items (80% machine-scored) in each of the Core testlets. Within each unit, there needed to be as many uto-scored items as possible so that there was enough information to make routing decisions for the next. Human-coded items could not be scored on-the-fly, thus, were not used for routing decisions. Thus, the five units assigned to core testlets had to be those that had the highest number of machine-scored items and included items of average difficulty. It was also important to represent both trend and new units as equally as possible across the core testlets.
- Stage 1: Each testlet in Stage 1 consisted of three units with approximately 15 items per testlet and, like the Core stage, there needed to have a high proportion (60%) of auto-scored items so that most items could contribute to the routing decisions for Stage 2. As with the Core, it was important to balance the representation of trend and new units, so at least one trend and at least one new unit was assigned to each testlet. In stage 1, the test developers felt it was important to represent the new multiple source and/or digital reading framework elements within every testlet. This would ensure that every student, no matter what their skill level was, would have an opportunity to interact with the new aspects of the framework, primarily multiple sources. Thus, unit assignment within Stage 1 required balancing the number of items in a unit, the number of auto-scored items in a unit, the trend or new status of a unit, whether the unit contained a single source or multiple sources and the average difficulty of the unit (low or high average difficulty).
- Stage 2: It included the longest units both in terms of the number of items and the length of texts. Assignment of units in this stage could also contain a higher proportion of human-coded items because there were no more routing decisions in this stage in the original design (Design A). In this stage, testlets were constructed such that each testlet had a wider range of difficulties. Here, each testlet had only two units and approximately 14 items total, across those units. Every effort was made to include at least one new unit that represented the new aspects of the framework as was done in the stage 1 testlets.

Review by the Reading Expert Group

Once the item selection was complete and the units were assigned to the multistage adaptive design, the psychometricians performed simulation studies to assess the performance of the design using the preliminary item parameters obtained from the field trial. The details of these simulation studies are described in Yamamoto, Shin, & Khorramdel (2019). In short, the simulation studies suggested that the item parameters could be recovered well with minimal errors and that the proposed multistage adaptive design (i.e., combination of Design A for 75% of students and Design B for 25% of students) would improve the measurement precision for all range of skill distribution, particularly at the extreme lower and higher ends of distribution.

More specifically, the simulation study showed about 5% reduction of measurement error on average over the range of typical skill distributions, and 11% more test information compared to the similar number of randomly selected units.

Given that the multistage adaptive testing design consisted of 192 possible paths, it was not possible for the reading experts to review all those combination of item sets and make recommendations for the selection. Instead, at the REG meeting following the field trial, a thorough explanation of the item selection process and the unit assignment was presented as well as the characteristics of the final item pool. The item pool was evaluated at a holistic level, considering the representation of the cognitive processes across the entire pool and the distribution of difficulty and construct representation within each stage of the multistage adaptive design. At the end of the meeting, the experts signed off on the multistage adaptive design and the unit assignment.

Reading Fluency

The field trial data for the reading fluency items was reviewed to make sure all items were functioning as intended across country-by-language versions. While there was a range of difficulty of the items which was a little greater than expected, the decision was made to keep all 65 items for the main survey. The advantage was that all testlets could remain the same, resulting in the same design used in the field trial.

Global Competence

Item selection for global competence followed the same approach used by reading whereby classical item statistics were first used to eliminate rows of complex-multiple choice, table items. Scoring rules were developed for the table items. Then, in a second phase, IRT analyses were conducted and IRT results was used to eliminate items that did not perform well. From the original 86 items tested in a separate pilot study, two were combined because of inter-item dependencies, 17 items were dropped from the MS item pool and a new item developed for the MS per request of the expert group. All of the units tested in the field trial were retained for the main survey which test developers felt was a strong benefit to the selection. Maintaining the same range of contexts from the field trial to the main survey provided good continuity and kept the range of topics broad.

Clusters were created following the final item selection and balanced based on the coverage of cognitive processes, the discrimination and difficulty of the items, the total number of units and items, the number of constructed-response items and the content covered in the topic of each unit such that similar topics did not appear in the same cluster. IRT test information curves were evaluated across clusters to ensure that they were balanced.

The global competence experts reviewed the pilot study data, the approach to item selection, the content and balance of the clusters and signed off on the selection. The expert group requested adding a new item to one of the existing units. This was done through a review process between the test developers and the experts after the meeting, and once it was finalized, it was added to the item pool for a total of 86 items.

Financial Literacy

Like reading and global competence, the item selection for financial literacy followed several stages. First, classical item analysis was used to eliminate poorly functioning rows in the complex-multiple choice table items. Scoring rules were then created for each complex-multiple choice. Then, IRT analyses were conducted and the IRT results were used to eliminate items that did not function well. Six new items were eliminated based on the item performance data.

A final stage of item selection eliminated some trend items and two new clusters were created which combined the trend and the new items. This selection included dropping the trend items that did not function well and were outdated. A total of 43 items were selected from the original item pool of 63. Two clusters were created by balancing the financial literacy items based on construct coverage, number of items and length.

The item counts for reading, mathematics, science, global competence, and financial literacy in both the field trial and main survey are presented in Table 2.7.

Table 2.7 Item counts (field trial and main survey) by domain and delivery mode

Construct coverage

The set of items for the main survey was balanced in terms of construct representation, based on the overall distributions recommended in the frameworks.

A total of 245 items – 72 trend and 173 new items –were selected for the computer-based multistage adaptive design for reading, with the distribution as shown in Table 2.8 below. Of the 173 new items retained for the main survey, 54% were originally submitted by countries, 20% were originally drafted at the Item Development Workshops and 26% were created by test developers at ETS.

Table 2.8 Reading item counts by framework category (CBA)

The 69 items selected in the global competence domain were distributed among the framework categories as shown below in Table 2.9 and Table 2.10.

Table 2.9 Global competence item counts by framework category: Cognitive processes

Table 2.10 Global competence item counts by framework category: Cognitive subprocesses

Error! Not a valid bookmark self-reference. and Table 2.12 show the distributions of the 43 financial literacy items across the two aspects of the framework: process and content.

Table 2.11 Financial literacy item counts by framework category: Cognitive processes

Table 2.12 Financial literacy item counts by framework category: Cognitive content

Preparation of data collection instruments

Preparing the main survey national student delivery systems (SDS)

The process for creating the main survey national student delivery system (SDS) followed the approach used during the field trial, beginning with assembly and testing of the master SDS followed by the process for assembling national versions of the main survey SDS.

After all components of national materials were locked, including the questionnaires and cognitive instruments, the student delivery system was assembled and tested first by Core 2. Countries were then asked to check their SDS and identify any remaining content or layout issues. Once countries signed off on their national SDS, their final systems were released for the field trial.

Preparing paper-based instruments

Like the field trial, national versions of the paper-based trend clusters were prepared by the Core A contractor. To better ensure comparability of the paper-based assessment materials across countries and languages, booklets were centrally created by Core A and then reviewed and approved by countries.

The first step was to review and finalized clusters and common booklet parts (i.e., cover page, introduction). The approved clusters were then assembled into the 30-min survey 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.

Main survey coding

The process used for the main survey coding training was slightly different from that employed prior to the field trial as it included full training for all main survey items, both new and trend items. 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 students' responses. Queries were reviewed, and responses were provided by domain-specific teams including item developers and members of the response team from previous cycles.

Revisions were made to the coding guides for reading and global competence following the field trial and field trial pilot, respectively. The coder queries helped test developers see response categories that weren't anticipated during the development of the coding guide. Thus, based on the queries received, test developers made some coding guides clearer and added sample responses to the guides to better illustrate different types of responses. Workshop examples were also enhanced to include more authentic student responses to better illustrate boundaries between full credit, partial credit (where appropriate) and no credit. Following the international coder training, additional revisions were made to the reading, global competence and financial literacy coding guides for the new items in response to discussions that took place at the meeting.

Released items to illustrate the framework

As has been the case in previous PISA cycles, several items were released to the public domain at the time of publication of the PISA 2018 results to illustrate the kinds of items included in the assessment. The OECD decided to release a reading unit from the main survey: *Rapa Nui* (7 items). In addition, the OECD decided to release one unit from the initial item pool prior to the field trial –*Galapagos* (7 items) – and two units from the field trial pool – *Chicken Forum* (7 items), and *Cow's Milk* (7 items). These units are available at *www.oecd.org/pisa*.

REFERENCES

Curran, P. J. et al. (2008), ”Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis”, *Developmental Psychology*, Vol. 44/2, pp.365-380, *http://doi.org/10.1037/0012-1649.44.2.365*.

Glas, C. and K. Jehangir (2013), “Modeling country-specific differential item functioning”, in L. Rutkowski, M. von Davier and D. Rutkowske (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, CRC Press, Taylor and Francis Group, Boca Raton, FL.

Glas, C. A. W. and N.D. Verhelst (1995), “Testing the Rasch Model”, in G. H. Fisher and I. W. Molenaar (eds.), *Rasch models: Foundation, recent developments, and applications*, pp. 69-96, Springer-Verlag, New York.

Mazzeo, J., and M. von Davier (2013), “Linking scales in international large-scale assessments”, in L. Rutkowski, M. von Davier and D. Rutkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, CRC Press, Taylor and Francis Group, Boca Raton, FL.

Meredith, W. and J.A. Teresi (2006), “An essay on measurement and factorial invariance”, *Medical Care*, Vol. 44/11, pp. 69-77.

Meredith, W. (1993), Measurement invariance, factor analysis, and factorial invariance, *Psychometrika*, Vol. 58/4, pp. 525-543.

Oliveri, M. E., and M. von Davier (2014), “Toward increasing fairness in score scale calibrations employed in international large-scale assessments”, *International Journal of Testing*, Vol. 14/1, pp. 1-21.

Oliveri, M. E., and M. von Davier (2011), “Investigation of model fit and score scale comparability in international assessments”, *Psychological Test and Assessment Modeling*, Vol. 53/3, pp. 315-333.

Reise, S. P., K. F. Widaman and R. H. Pugh (1993), “Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance”, *Psychological Bulletin*, Vol. 114/3, pp. 552-566.

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M., (2019). *Evaluating item response theory linking and model fit for data from PISA 2000–2012. Assessment in Education: Principles, Policy, and Practice, 26*(4), 466-488. doi:10.1080/0969594X.2019.1586642

Yamamoto, K, Shin H. J., & Khorramdel, L. (in press; 2018)

Appendix 1: Overview of the adaptive process for the standard design – Design A – that connects Core>Stage 1>Stage 2 (64 paths in total, applicable to 75% of students)

Appendix 2: Overview of the adaptive process for the alternative design – Design B – that connects Core>Stage 2>Stage 1 (128 paths in total, applicable to 25% of students)