PISA 2022 Technical Report

11 Scaling PISA data

Overview

The test design for PISA 2022 follows the balanced incomplete block (BIB) design used in prior cycles, with adaptations to incorporate multi-stage adaptive testing (MSAT) for the reading and mathematics domains. With the traditional BIB design, units (i.e., small sets of items) are grouped into mutually exclusive clusters (i.e., sets of units) assembled into test forms. For the non-adaptive domains, the clusters are distributed so that they appear with equal frequency across forms and positions within forms, which leads to the design being balanced. When these tests are administered, students are administered a randomly selected test form so that differences in the average test performance on forms consisting of different sets of items are not due to differences in student proficiency. However, the test forms can be of different difficulty, which means that the performance of groups measured through different sets of items cannot be directly compared using total-score statistics such as the average number or percent of items that the student responded to correctly.

The limitations of using the number or percent of items correct to score assessments that are designed with BIB or administered through MSAT can be overcome by modelling the item responses through item response theory (IRT). When students respond to a set of items in a common subject or domain, their response patterns should show regularities that can be modelled using the underlying commonalities among the items. This regularity can be used to characterize the students and items on a common scale, even when students take different sets of items. However, IRT is only the first step in the scaling of PISA data that makes it possible to describe the distributions of student performance in populations or subpopulations, to estimate the relationships between proficiency and background variables, and to build and select test forms that match the difficulty of the form with the ability of students.

The scaling approach employed in the analyses of PISA data (*population modelling*) combines IRT and latent regression modelling to increase overall measurement accuracy and to avoid potential bias in the estimation of the relationships between proficiency and contextual variables from the background questionnaire (BQ). Once the population model is estimated, multiple plausible values can be drawn for each student from a posterior distribution of proficiency that accounts for the sources of uncertainty in the data.

In PISA 2022, mathematics and reading MSAT designs were incorporated into the overall BIB design to deliver a 60-minute MSAT to students, instead of the two 30-minute clusters used for the other domains. The reading design was the same that was used in 2018. However, as reading became a minor domain, some of the items were released and the 2018 testlets that lost some items were re-assembled from the reduced item pool in a way that minimized the changes. As in 2018, the reading design included a proportion of student misrouted from the core to stage and from stage 1 to stage 2 to ensure that responses on all items were collected from students across a broad proficiency range. The reading design partially balanced item position between stage 1 and stage 2. For mathematics, a newer design was implemented that fully balanced item position across core, stage 1 and stage 2 and randomly assigned 25% of the students to a linear design to ensure that item responses are collected from students across a broad proficiency range (for further details, see Chapter 2 in this report).

However, despite these design differences across domains, for the most part, the same classical analysis (item analysis - IA and timing), item response theory (IRT) and population modelling procedures could and were effectively implemented to fulfil all the main survey analyses goals.

This chapter first describes the quantity and quality of the data submitted by the participating countries/economies. Analyses were conducted to evaluate how well the assessment design was reflected in the data and to verify that the data quality was appropriate for IRT and population modelling. The subsequent sections explain the models and methods used for IRT, latent regression modelling, and the generation of plausible values. Then, the application of these models and methods to the PISA 2022 data to produce the national and international item parameters and the plausible values are described. Finally, the approach and methods used for estimating the linking errors between the 2022 main survey and the previous PISA cycles are explained.

Data yield and data quality

Before the data were used for scaling and population modelling, analyses were carried out to examine the quality of the data to ensure that the test design requirements were met, and also to verify that the data reflected the intended design. The following subsections give an overview of these analyses and their results. Overall, the quality of the data and the cognitive instruments met the requirements for the intended analyses and scaling methods. The results of the item analyses were communicated to countries/economies for their review and feedback. Taken together, the data yield and item analyses confirmed that the PISA 2022 computer platform had successfully delivered, captured, and exported the student- and item-level data expected from both the computer-based assessment (CBA) and paper-based assessment (PBA).

Target sample size, routing, and data yield

Target sample size

The assessment design for the PISA 2022 main survey included the core domains of reading, mathematics, and science, delivered through both CBA and PBA. In addition, it also included the optional domain of financial literacy and the innovative domain of creative thinking, both delivered only through CBA. As part of the sampling design, participating countries/economies were required to sample a minimum of 150 schools to cover their national population of 15-year-old students. Countries/economies taking the CBA with creative thinking (CrT) or the CBA without CrT needed to sample 42 students from each of the 150 schools for a total sample of 6,300 students, while countries/economies taking the PBA needed to sample 35 students from each of the 150 schools for a total sample of 5,250 students. CBA countries/economies taking the financial literacy domain were also required to sample more schools and/or more students per school to obtain an additional sample of 1,650 students, resulting in a total sample of 7,950 students. This group of 1,650 students who took the financial literacy sample was randomly equivalent to, albeit different from, the “main sample” students who did not take financial literacy.

With mathematics as the major domain, one hour of mathematics was administered to most of the students in the main sample (i.e., 96% with CrT and 94% without CrT), and the other domains were only administered to a subset of students.

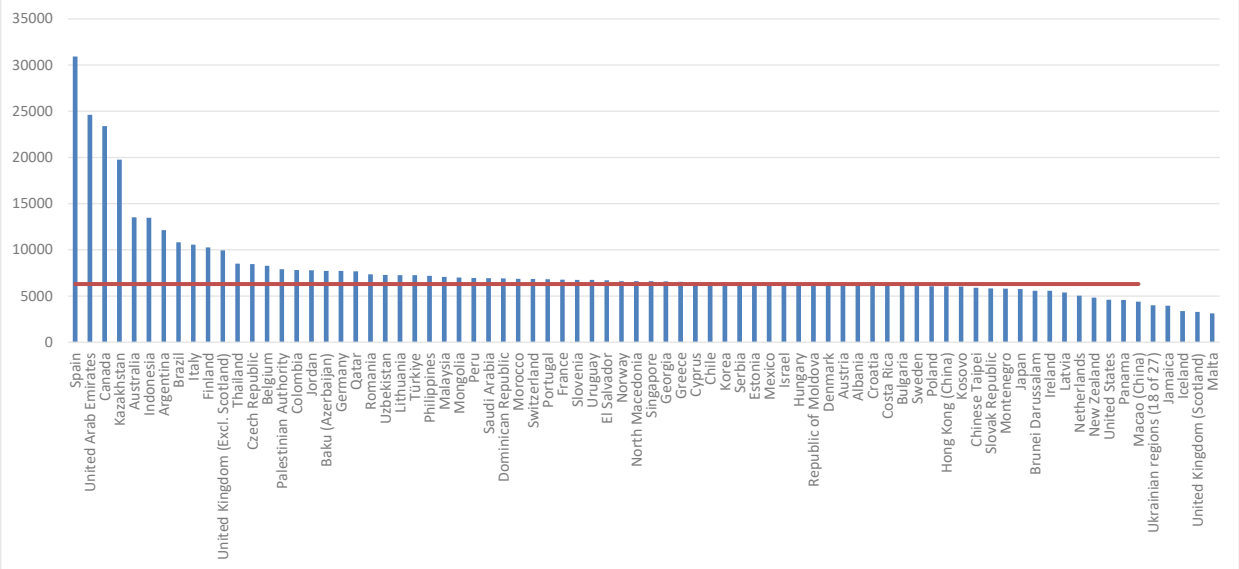
Data yield

Table 11.1 shows the assessment languages and the sample sizes for each of the participating countries/economies. For a student to be considered a “respondent” for PISA, the student needed to meet at least one of the following two criteria: 1) answered more than half of the cognitive items from the

assigned form/booklet, or 2) answered at least one cognitive item and at least one item regarding home possessions (i.e., ST251 or ST255).

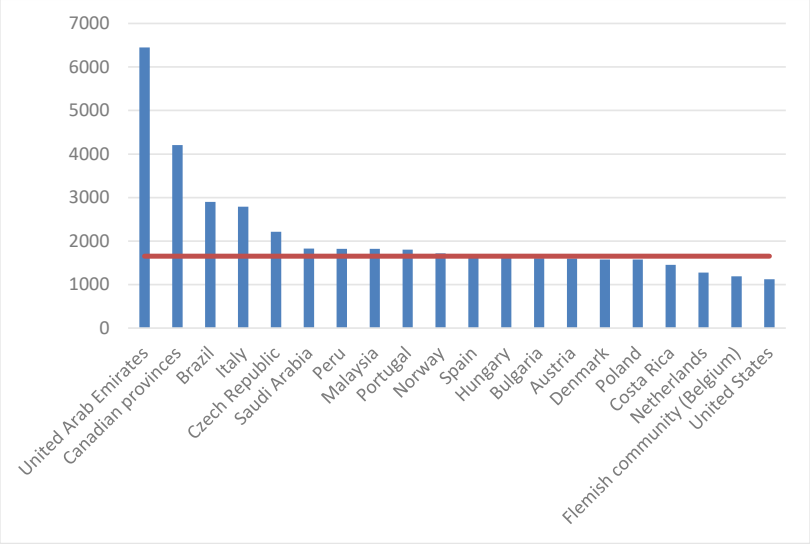
Figure 11.1, Figure 11.2, and Figure 11.3 show the extent to which each country/economy participating in the CBA, the financial literacy assessment, and the PBA met or exceeded the sample size requirements. In each figure, the red horizontal line indicates the sample-size requirements for each design option. Some countries/economies exceeded the requirements because they oversampled certain regions and/or minority languages. As expected, a few countries/economies did not reach the sample size requirements because of their small total population size. Because of on-going post-Covid challenges, 26 countries/economies did not reach their sample-size target. Nevertheless, most of them managed to get very close, and all collected enough data to contribute to the international scaling and to produce high-quality population modelling outcomes that are comparable to those of all other participating countries/economies.

Figure 11.1. Main sample yield for countries/economies participating in the CBA



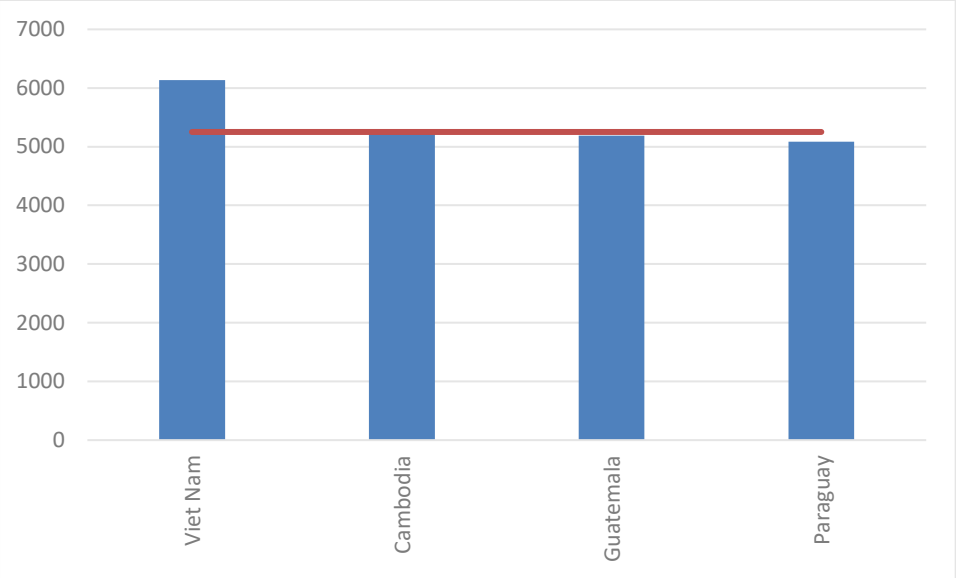
Note: Ukranian regions (18 out of 27) administered the assessment.

Figure 11.2. Financial literacy sample yield for participating countries/economies



Note: 'Canadian provinces' refer to the seven provinces of Canada that participated in the PISA 2022 financial literacy assessment: British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario and Prince Edward Island. It is not a nationally-representative sample. 'Flemish community (Belgium)' refers to the Flemish-speaking population of Belgium. It is not a nationally-representative sample.

Figure 11.3. Main sample yield for countries/economies participating in the PBA and new PBA



Since the sample sizes varied greatly across countries/economies, the number of sampled schools and the sample sizes from each school varied as well. As shown in Table 11.1, the number of schools ranged from 46 to 983, but most countries/economies met the requirement to sample a minimum of 150 schools.

The PISA 2022 assessment design also required that students be randomly assigned to forms in the prescribed proportions. Results showed that this condition was met for all participating countries/economies and that the assignment of students to items was appropriate for the item analyses and IRT scaling.

MSAT data yield for mathematics and reading

The goal of the mathematics and reading MSAT designs was to improve the measurement precision across a wide range of proficiencies, and at the same time, to collect optimal data needed for the item analyses and IRT scaling. Therefore, it was important to verify that the MSAT design was implemented as intended. Note that some students in some countries/economies took a shorter non-adaptive *U*ne-heure (UH) booklet/form. Also, in Israel, some students took a non-adaptive *Ultra-Orthodox* (UO) form. These UH and UO cases were excluded from the MSAT analyses reported in this chapter.

Four critical aspects of the MSAT designs were closely monitored:

- Random assignment to each routing testlet
- Random assignment to each of the adaptive (75%) or linear MSAT paths (25%) in mathematics
- Random assignment to each of the Design A (75%) or Design B (25%) paths and misrouting in the expected proportions, in reading
- Adaptive second and third testlet selection according to students' observed performance on prior stages according the MSAT design.

Near uniform proportions of the total number of alternatives were observed that confirmed the random assignments to the routing testlets, the alternate MSATs (Groups A, B and C in mathematics and Designs A and B in reading[1]), and the mathematics adaptive and linear paths.

The adaptive routing through the mathematics and reading designs are summarized in Figure 11.4a and Figure 11.4b, showing the proportion of students in each country/economy who were routed to difficult, medium or easy testlet combinations such as: hard testlets in both Stage 1 and Stage 2 in mathematics or reading (HH); or low and medium or hard and medium testlet combinations in mathematics (LM or HM), or low and hard or hard and low difficulty testlets combinations in reading (LH or HL); or low difficulty testlets in both Stage 1 and Stage 2 in mathematics or reading (LL). Students' paths were categorized as missing/undetermined when they did not complete the routing stage or stage 1 and their full path could not be determined by the adaptive algorithm. Note that the 25% of students who were assigned to non-adaptive paths in the hybrid MSAT design are not included in Figure 11.4a.

In both figures, the lowest to highest performing countries/economies are shown from left to right. As intended by design, in the lower-performing countries/economies, a smaller proportion of students were assigned to the most difficult testlets, while in the higher-performing countries/economies, a smaller proportion of students were assigned to the easiest testlets. Also, as intended, every type of testlet was assigned to a high enough proportion of the total sample in each country/economy in each stage, regardless of the proficiency distribution in the country/economy. For reading this was achieved through the misrouting of some students, while for mathematics this was achieved by randomly assigning 25% of students to non-adaptive paths of the hybrid MSAT design. Altogether the observed results confirmed that the MSAT delivery platform worked as intended, and that regardless of the countries/economies' proficiency distributions, the adaptive design always provided the minimum number of responses per item needed for IRT scaling and an appropriate item coverage across the full range of student proficiency.

Figure 11.4a. Proportion of students routed to each testlet combination in mathematics MSAT

Refer to Chapter_11_Figures.xlsx to view this figure on line.

Figure 11.4b. Proportion of students routed to each testlet combination in reading MSAT

Refer to Chapter_11_Figures.xlsx to view this figure on line.

Classical test theory statistics: Item analysis

Classical item analyses (IA) were conducted on all paper-based and computer-based test items at the national and international levels to verify that the items functioned appropriately. Unexpected results were identified and explored for any indication of possible issues related to data collection, human- or machine-scoring, or other issues. Descriptive statistics for the observed responses and various missing response codes were provided to countries/economies and the OECD for their review and feedback. Classical item analysis also provided additional descriptive information useful for the review of the IRT modelling outcomes.

The following statistics were computed:

- item response category statistics, including frequency and criterion score mean, standard deviation, and biserial correlation
- (classical) item difficulty
- (classical) item discrimination

Item response categories included several types of non-response and item score categories. An item response was recoded as *not-reached* when a student did not answer the item or any subsequent item in

the cluster for non-adaptive domains (science, financial literacy, and creative thinking) or in the MSAT sessions for reading and mathematics. An item response that did not perform properly in the field or had a missing human-coded response code was also converted to not-reached. An item response was recoded as *omitted* when a student did not answer the item but answered one or more of the subsequent items in the cluster or the MSAT path. The category *off-task* was used to identify an invalid missing category when a student did not answer the question in the expected way (e.g., by giving a response not associated with the item or responding with more than one answer in an exclusive choice question). In the computation of the item statistics and in the scaling analyses, the not-reached responses were excluded (i.e., treated as missing/ not-administered), but the omitted and off-task responses were treated as incorrect.

The mean score, standard deviation, biserial/polyserial correlation, and point biserial/polyserial correlation were based on the total block/cluster score where the item appeared.

Statistics for trend items were compared with results from prior PISA cycles. Also, statistics were compiled separately for the PBA and CBA and were examined at the aggregate level across countries/economies. Analyses were also performed separately for each country/economy to identify outlier items that worked poorly or differently across assessment cycles and/or across countries/economies and to detect flaws or obvious scoring rule deviations. Analyses were also conducted by language within each country/economy. UH booklet results were provided for countries/economies, where applicable.

Table 11.2 and Table 11.3 show examples of the item analysis outputs. Table 11.2 shows the IAs of the first three items in block/cluster M01 of one country/economy. The first item, DM033Q01C, is the scored version of the paper-based item PM033Q01 (the corresponding CBA item is CM033Q01), a multiple-choice item. Each section of the table represents one item, and the columns represent the different response categories. The *total* column includes the summary information for all categories, excluding the not-reached (*NOT RCH*) category. The last row (*RSP WT*) shows the scores associated with each response category and the maximum score that can be obtained on the item.

The biserial (*R BIS)* statistic is used to describe the relationship between performance on a single test item and a criterion (usually the total score on the test). It is estimated using the polyserial method which is a generalized form of the correlation between the criterion (which is treated as a continuous variable) and the item score, where the item score is either 0, 1 (for dichotomous items) or 0, 1, 2, 3,…, *k* (for polytomous items).

The delta statistic is an index of item difficulty based on P+ (proportion correct, or percent correct when expressed as a percentage) which has been transformed so that it is on a scale with a mean of 13.0 and a standard deviation of 4.0. Delta statistics ordinarily range from 6.0 for a very easy item (approximately 95% correct) to 20.0 for a very difficult item (approximately 5% correct), with a delta of 13.0 corresponding to 50% correct.

Table 11.3 has two parts. The first part shows a breakdown of the score categories and biserial correlations by category. The second part contains summary data for each item and reveals items that were flagged for surpassing certain thresholds. The thresholds are provided in Table 11.4. In this example, the third item is flagged for having an omit rate of greater than 10%.

Response time analyses

The computer-based platform captured response time data for all computer-based items delivered in the CBA countries/economies in both the field trial and main survey. Timing data can be informative in evaluating the level of student engagement and effort over the two-hour testing period. Very little time spent on the assessment was interpreted as low effort, while too much time spent on the assessment (or parts of the assessment) could be an indication of technical problems or low ability. Response time information was aggregated by testlet, cluster, domain, and for the full assessment. Item response times by position and proficiency level were also computed. Overall, results indicate that the CBA data provided

valid information that can be used to model items and estimate student performance within and across countries/economies.

Outliers

Students were generally expected to complete the cognitive assessment within two one-hour periods separated by a break. Within each hour, students followed the prescribed order of clusters or MSAT testlets and units at their own pace. Except for the CBA reading and mathematics assessments, students were expected to complete two 30-minute clusters within an hour, regardless of the positions within the assessment (e.g., clusters 1 and 2 in the first hour, clusters 3 and 4 in the second hour). Within each hour, students were allowed to manage their time between the two assigned clusters. For reading, students were expected to complete the reading fluency items within a 3-minute limit and three self-paced MSAT routing, stage 1 and stage 2 testlets (i.e., testlet 1, 2 and 3) within the remaining time in the hour. For mathematics, students were expected to complete three self-paced MSAT or linear testlets within the hour.

Focusing on larger-than-expected cluster or testlet response times, outliers were identified using the median absolute deviation (MAD) approach (Leys et al., 2013[1]; Rousseeuw and Croux, 1993[2]). That is, response times greater than median$\{x_i\}$ + 4.4478*median$\{|x_i - \text{median}(x_j)|\}$, where $\{x_i\}$ is the collection of all sample values and $|\cdot|$ denotes their absolute value, were identified as outliers. Note that in this calculation, median values were identified using international data, not country/economy-level data. This way, the same criterion was used across countries/economies, and the identification of outliers was more stable.

Table 11.5 shows the percentages of response time outliers by domain. The proportions of outliers were small—between 0.5% to 1.2% across all domains. Note that, because reading fluency was very short and strictly time-limited, an outlier analysis was not needed.

Cluster- or testlet-level response time

Table 11.6a presents descriptive statistics for testlet or cluster response times for all CBA domains, excluding reading fluency. These values are the sum of the time each student spent on each item in a testlet or cluster, aggregated across students, countries/economies, and positions. Similarly, Table 11.6b presents descriptive statistics for domain time, computed as the aggregated item time.

These results show that most students spent a reasonable amount of time on each cluster (with most taking more than 13 minutes and less 30 minutes, approximately from the first (Q1) to the third quartiles (Q3)) or on each testlet (more than 13 minutes and less 30 minutes, approximately Q1 and Q3) or each testlet (more than 6 minutes and less than 22 minutes, approximately Q1 and Q3). However, as sample maximum (MAX) values show that some students did take a large amount of time to complete a given 30-minute cluster, thus and having very little time to finish the subsequent cluster with which it was paired. Similarly, for mathematics and reading, values show that some students did take a large amount of time to complete the first or the first two testlets and have little time for the subsequent(s) one(s). It is also notable that the last mathematics and reading testlets generally took less time than the other testlets.

Total domain time was also appropriate in all domains, with most students spending more than 30 minutes (Q1) and less than 54 minutes (Q3). Overall, the time spent in each domain was quite similar, although science and financial has larger Q3 and MAX values. Also, a desired confirmation was that there was no evidence of a timing mode effect between the linear and MSAT groups in mathematics and between design A and B in reading.

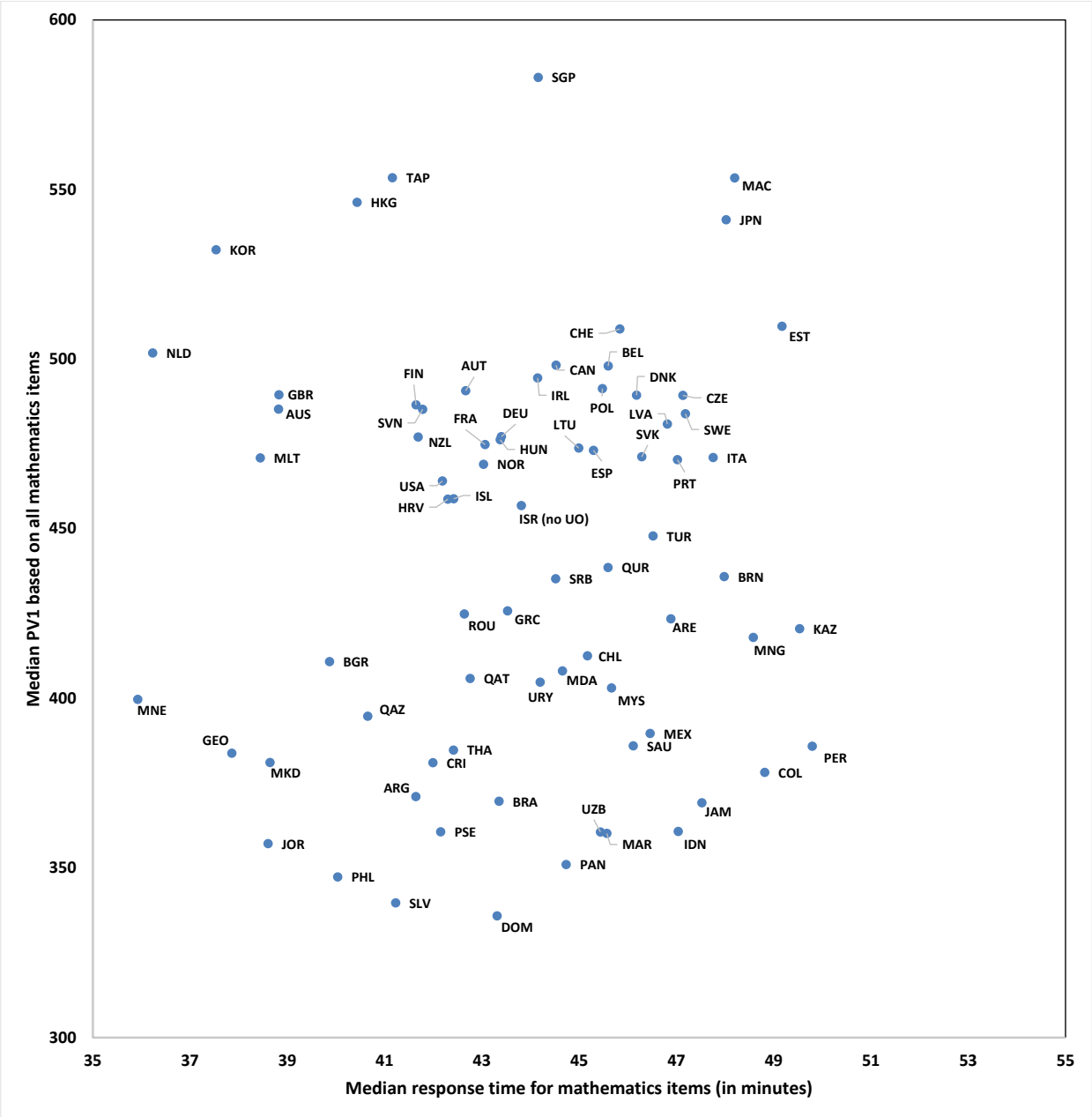
Response time and student performance

The relationship between response time and student performance was examined using the median of the cluster-level response time and proficiency levels. The proficiency levels were computed based on the first plausible value (PV1) and a detailed description of their interpretation and cut-offs can be found in Chapter 17. Tables 11.7a – 11.7d show a very similar pattern across all domains and MSAT designs, where from Below Level 1 and up to Level 4, more able students generally spent more time completing

each domain. The increase in time spent was most noticeable between students below Level 1 and up to Level 3; then, time spent tapered off up to Level 5 and slightly decreased at Level 6. Again, there was very little difference between the linear and MSAT mathematics tests, except at Levels 5 and 6 where the MSAT students spent about one to two minutes more in median time than the linear students.

While the more proficient students generally took more time to complete the test, median time and median performance varied noticeably across countries/economies. However, as Figure 11.5 shows for mathematics, while countries/economies do vary noticeably in their median PV1 proficiency, there was no clear relationship between median proficiency and median total item response time across at the country/economy level. For example, KOR and SGP, both have high median mathematics scores, but SGP's median response time is close to the overall median response time, while KOR's is well below it.

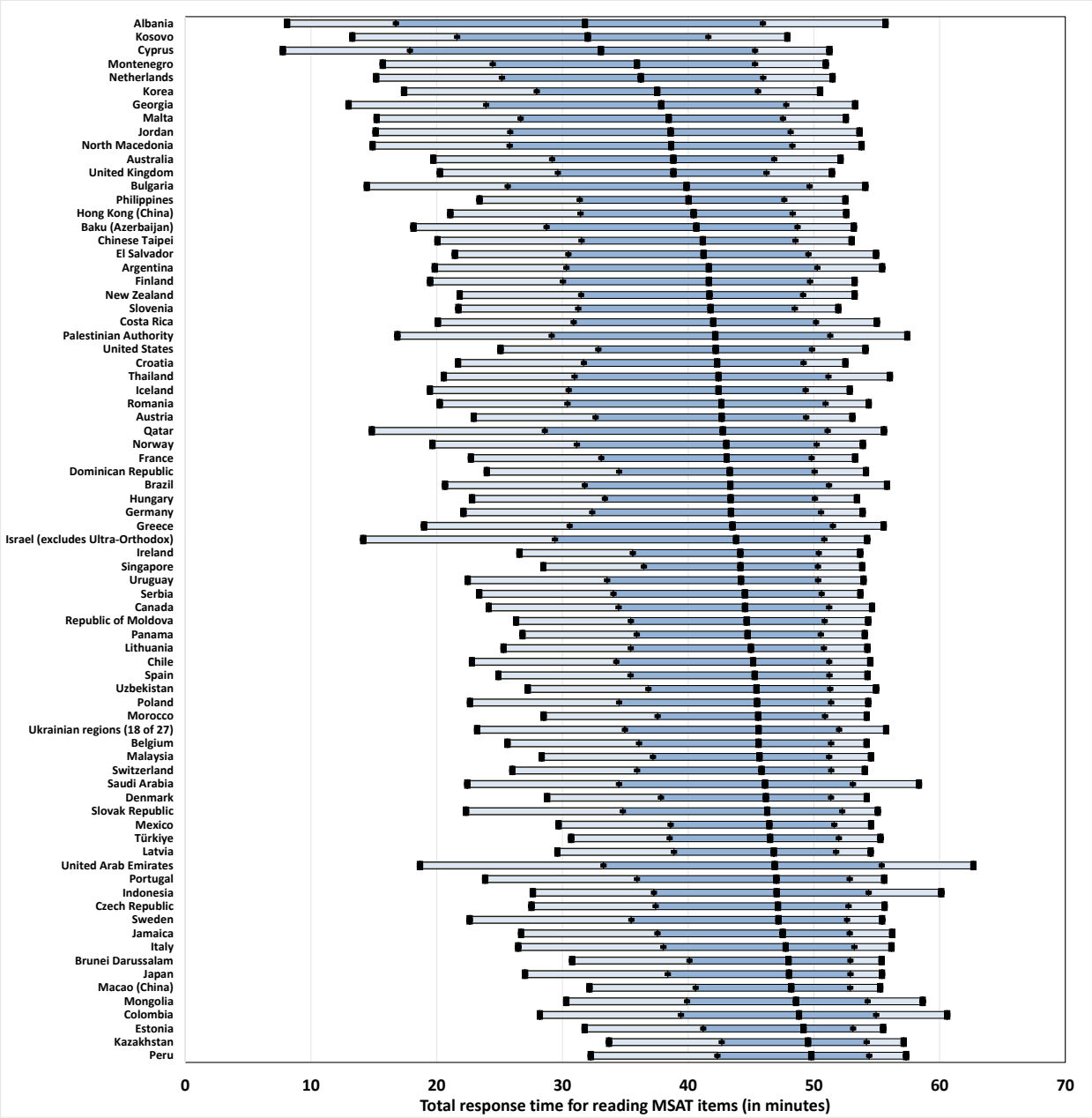
Figure 11.5. Mathematics median response time by median proficiency across countries/economies



Note: Statistics were calculated only with students who had timing data, excluding UH students. For Israel, Utra Orthodox students were excluded.

Because of differences in proficiency and other factors, the time it takes students to complete the assessment is expected to vary within each country/economy. This is shown in Figure 11.6 which presents the distribution of the total time spent on the mathematics items for all countries/economies, sorted by the median response time. Note that in a few cases the 90th percentile time was above 60 minutes allocated. This was because the time limit was not strictly enforced to allow for students to finish tasks they were in the middle of.

Figure 11.6. Distribution of mathematics response time in each country/economy



Note: For each country/economy, the solid black line in the middle shows the median total response time, the dark blue horizontal bars range from the 25th to the 75th percentiles, and the light blue horizontal bars range from the 10th to the 90th percentile. Countries/economies are sorted by their median MSAT response time. For Israel, Utra Orthodox students were excluded.

Item-level response time

Response time and the relationship between response time and performance were also explored at the item level. Figure 11.7 and Figure 11.8 show the median item-level response time (aggregated across all countries/economies) for the trend and new mathematics items, respectively, disaggregated by students' proficiency levels based on PV1. For most but not all items, as we have seen above with the total domain time, low-performing students (blue and red lines) had similar and relatively short response times, while high-performing students (green and purple lines) had longer response times and larger variability in the response times. This pattern was consistently observed for both the trend and new mathematics items. Furthermore, there are some clear peaks indicating items on which high-performing students spend substantial more time than low-performing students.

Figure 11.7. Median item response time by proficiency level for mathematics trend items

Refer to Chapter_11_Figures.xlsx to view this figure online.

Figure 11.8. Median item response time by proficiency level for mathematics new items

Refer to Chapter_11_Figures.xlsx to view this figure online.

For the creative thinking items, median item response times are shown in Figure 11.9, for each country/economy. A similar approach was employed but levels were calculated using the first non-linear score transformed value instead of PV1. More detail on the CrT scores and their levels can be found in Chapter 18. As expected, since items were typically more demanding and fewer of them were administered, students generally spent more time per item than for the other domains. Across countries/economies, the amount of time spent per item varied, however, the timing patterns across items were similar.

Figure 11.9. Median item response times for creative thinking items

Refer to Chapter_11_Figures.xlsx to view this figure online.

Response time reflecting possible motivation or administration issues

On average, students completed the entire test in 83.34 minutes (excluding a short break between the two assessment hours), with a standard deviation of 21.74 and a median of 87.46 minutes. Some students completed the test in less than 30 minutes (found in all countries/economies, 2.7% of the overall sample), while some students took longer than 120 minutes to complete the test (1.5% of the overall sample). At the country/economy-level, students in Kazakhstan, Peru, Mongolia, and Macao took the longest time to complete the entire test, with a median time of 100.8, 99.0, 99.0 and 98.9 minutes, respectively. Students in Cyprus, Albania, and Kosovo took the shortest time to complete the test, with a median time of 67.2, 67.7 and 69.5 minutes, respectively.

There were five countries/economies where 5% or more of the students exceeded the time limit: United Arab Emirates (11.9%), Indonesia (9.9%), Colombia (7.6%), Mongolia (5.7%) and Saudi Arabia (5.4%). This could be explained by students in these countries generally spending more time to complete the test and by the fact that time limits were not strictly enforced so that students in the middle of a task could finish without being abruptly cut-off. Apart from these countries/economies, only a small proportion of respondents in each country/economy had very long or short total response times, indicating that there were no systematic administration and/or motivation issues. Furthermore, students with these extreme response times appeared to be randomly distributed across schools and countries/economies.

Position effects

According to the PISA test design, each student takes one of many alternative test forms made up of different clusters/testlets in different positions. For example, a student may take two science clusters in the first hour and then take three mathematics testlets in the second hour, while another student may take the same domains, but in the reverse order. Item position effects are a concern in large-scale assessments because substantial position effects, if present, would increase measurement error and may introduce bias in parameter estimation. To mitigate any potential item position effects, as in previous cycles, the PISA 2022 main survey design balanced the order of the domains (between the first and second hour) as well as the order of the clusters or testlets within each domain (see Figure 2.5 in Chapter 2 for the full form design used in PISA 2022). Thus, PBA and CBA clusters and items within them (in fixed position) appeared in the first hour in positions 1 and 2 and in the second hour in positions 3 and 4. The CBA testlets for mathematics appeared in the first hour in positions 1, 2, and 3, and in the second hour in positions 4, 5, and 6. The exception was reading, where the MSAT design was partially balanced with the core testlets appearing in positions 1 and 4 and the stage 1 and stage 2 testlets each appearing in positions 2, 3, 5, and 6.

As prior PISA cycle results have indicated, the PISA 2022 results summarized below show that position effects are significant and justify the use of the complex BIB and balanced MSAT designs implemented to minimize their impact.

To evaluate and confirm that the impact of item positions studied in the field trial was minimal in the PISA 2022 main survey, position effects were examined in terms of: 1) proportion of correct responses, 2) median response time, and 3) rate of omitted responses. For PBA and CBA domains, cluster-level statistics are reported for positions 1, 2, 3, and 4, and position effects are reported as the difference between positions 4 and 1. For the mathematics and reading MSATs, domain-level statistics were reported for the 1st hour and the 2nd hour[2] and the position effects are reported by the difference between hour 2 and hour 1.

Table 11.8a and Table 11.8b present the position effects in terms of the median response time[3] averaged by cluster position and by assessment hour, respectively. For all domains, students spent more time on a cluster when presented in position 1 than in position 4. Financial literacy items had a noticeably higher median response time when in cluster position 1, resulting in a larger difference between the median response times for cluster positions 1 and 4. There were indications that some students spent much more time on clusters 1 and 3, leaving them with less time for clusters 2 and 4, respectively. Table 11.8b shows that the position effects by hour were generally smaller than the position effects by cluster. Across domains, students spent between 3.54 to 6.92 minutes less in median response time in the second hour. For mathematics, positions effects appear nearly identical between the linear and MSAT part of the hybrid design. For reading, the response-time position effect is larger for the core than the first and second stages.

Table 11.9a and Table 11.9b present the position effects in terms of the average P+, averaged by cluster position and by assessment hour, respectively. By cluster, the decreases in P+ between position 1 and 4 ranged from 0.051 in creative thinking to 0.89 in financial literacy. Overall, cluster position effects were similar to values observed in prior PISA cycles. By assessment hour (Table 11.8b), for all non-adaptive domains, a smaller decrease in P+ between the 1st and 2nd assessment hour was observed compared to the decrease in P+ between the 1st and 4th cluster position. For the mathematics linear and adaptive MSAT trend and new items, the decrease in average P+ between the 1st and 2nd hour were all relatively small and similar to the decreases observed in the other domains.

The proportions of omitted responses at different positions for all CBA countries/economies were analysed to further examine the quality of data affected by position. The proportion of omitted responses are shown by cluster position and assessment hour in Table 11.10a and Table 11.10b, respectively. These do not include the ‘not-reached’ items. Note that the proportion of omitted responses for reading fluency are 0

because students had to respond to each item presented (i.e., they were not able to skip the item). Overall, the omission rates by cluster and by hour were very similar across the domains. As in PISA 2018, the omission rates for all domains in all positions were less than 0.10, and the omission rates in positions 2 and 4 were higher than the rates in positions 1 and 3, respectively.

Position effects were also reviewed for the new PBA forms. Table 11.11a and Table 11.11b report the average P+ and the average omission rates by cluster position. By comparison with the results from the PBA forms used in the prior cycles, the new PBA position effect were noticeably smaller: Position 4 – Position 1 decrease in P+ by less than 0.04 (compared to less than 0.09) and Position 4 – Position 1 omits increased by less than 0.02 (compared to less than 0.05).

IRT modelling and scaling

The modelling and scaling of the PISA 2022 main survey data followed the general approach developed for PISA 2015 [OECD (2017[3]), Chapter 9]. The following sections describe the IRT models and their assumptions, as well as the IRT scaling approach used in PISA 2022. The scaling issues associated with the mathematics and reading MSAT designs and how they were resolved are addressed as well.

IRT models and assumptions

As in PISA 2015 and 2018, the unidimensional multiple-group IRT model (Bock and Zimowski, 1997[4]; von Davier and Yamamoto, 2004[5]) based on the two-parameter logistic model (2PLM) (Birnbaum, 1968[6]) for the binary item responses and the generalized partial credit model (GPCM) (Muraki, 1992[7]) for the polytomous item responses were used for each domain. The 2PLM is a generalization of the Rasch model (Rasch, 1960[8]), which assumes that the probability of a correct response to item *i* depends only on the difference between the student *v*'s trait level θ_v and the difficulty of the item b_i. In addition, the 2PLM postulates that for every item, the association between this difference and the response probability depends on an additional item discrimination parameter a_i:

Formula 11.1

$$P(x_{vi} = 1|\theta_v, b_i, a_i) = \frac{exp\ (Da_i(\theta_v - b_i))}{1 + exp\ (Da_i(\theta_v - b_i))}.$$

The probability of a positive response (e.g., solving an item correctly) is strictly monotonic, increasing with θ_v. The item discrimination parameter a_i, usually scaled by a constant D = 1.7, characterizes how quickly the probability of solving the item approaches 1.00 with increasing trait level θ_v when compared to other items. In other words, the model accounts for the possibility that responses to different items do not have the same weight with relation to the latent trait. The discrimination parameter a_i describes how well a certain item relates to the latent trait and, therefore, discriminates between examinees with different trait levels compared to other items on the test. One important special case of the model is when a_i = 1 for all items, in which case, the model is equivalent to a Rasch model.

The GPCM (Muraki, 1992[7]), like the 2PLM, is a mathematical model for the probability that an individual will respond in a certain response category on a particular item. While the 2PLM is suitable for items with only two response categories (dichotomous items), the GPCM can be used with items with more than two response categories (polytomous items). The GPCM reduces to the 2PLM when applied to dichotomous responses. For an item *i* with m_i + 1 ordered categories, the probability of obtaining a score of *k* (0, 1, 2,…, m_i) under the GPCM can be written as:

Formula 11.2

$$P(x_{vi} = k|\theta_v, b_i, a_i, d_i) = \frac{exp\{\sum_{r=0}^{k} Da_i\,(\theta_v - b_i + d_{ir})\}}{\sum_{u=0}^{m_i} exp\{\sum_{r=0}^{u} Da_i\,(\theta_v - b_i + d_{ir})\}},$$

where d_{ir} is the item-category threshold or step parameter as indicated in Appendix A), with $\sum_{r=1}^{m_i} d_{ir} = 0$ and $d_{i0} = 0$.[4]

Critical assumptions of most IRT models and the models used in PISA are conditional independence (sometimes referred to as local independence) and unidimensionality. Under conditional independence, item response probabilities depend only on the latent trait and the specified item parameters—there is no additional dependence on any demographic characteristics of the students, responses to any other items presented in a test, or the survey administration conditions. Under the unidimensionality assumption, a common single latent variable accounts for performance on the full set of items. With past PISA data, these assumptions have been verified and item parameters have been estimated for each cognitive domain separately through unidimensional IRT models. These assumptions need to be confirmed for each domain in which any new items are used.

With these assumptions, we can formulate the following joint probability of a particular response pattern $\boldsymbol{x}_v = (x_{v1},\ldots, x_{vn})$ across a set of *n* items:

Formula 11.3

$$P(\boldsymbol{x}_v|\theta_v, \boldsymbol{\beta}) = \prod_{i=1}^{n} P(x_{vi}|\theta_v, \boldsymbol{\beta}_i),$$

where $\boldsymbol{\beta}_i$ is the vector of parameters for item *i* from the associated IRT model. When replacing the hypothetical response pattern with the scored observed data, the above function can be viewed as a likelihood function that is to be maximised with respect to the item parameters. To do this, it is assumed that students (indexed *v=1, 2, …, N*) provide their answers independently of one another and that the student's proficiencies are sampled from a distribution $f(\theta)$. Using the sampling weights w_v, the likelihood function is, therefore, characterised as:

Formula 11.4

$$P(\boldsymbol{X}|\boldsymbol{\beta}) = \prod_{v=1}^{N} w_v \int P(\boldsymbol{x}_v|\theta, \boldsymbol{\beta}) f(\theta) d\theta.$$

Typically, the item parameters that provide the best possible fit to a given data set are estimated by maximising this function through a process called *item calibration*. The item parameters can then be used in the subsequent analyses, such as in the estimation of individual plausible values and population characteristics. However, it should be noted that IRT modelling does not provide an absolute scale, since any linear transformation of the item and latent trait parameters in the above formula leads to the exact same likelihood function, often referred to as scale indeterminacy or non-identifiability. Therefore, as part of the calibration process, a choice must be made for the IRT scale to be determined.

For further information regarding the IRT models discussed, see Fischer and Molenaar (1995[9]), van der Linden and Hambleton (1997[10]; 2016[11]), or von Davier and Sinharay (2014[12]) for the use of these models in the context of international comparative assessments.

IRT item calibration and scaling

The PISA data collection designs are complex, and the assessments are adapted and translated for each participating country/economy into one or more languages. To better account for potential cultural and language differences, and to optimally scale the item parameters and proficiency estimates across countries/economies and across modes (PBA and CBA), new calibration and scaling approaches were implemented in 2015. For each domain, a series of multi-group concurrent calibrations of the historical data (2015 and prior PISA cycles) were conducted (von Davier et al., 2019[13]) (OECD, 2017[3]), Chapter 9. As a result, all the items used in all the PISA cycles up to 2015 were estimated and scaled onto new common IRT scales (by domain) and new transformations from these IRT scales to the existing PISA reporting scale were established to preserved trend comparability.

For the first run of the series of multi-group concurrent calibrations, the item parameters were constrained so that only one set of *common* or *international parameters* was estimated per item to model the data for all the country-by-language-by-cycle groups. As part of the calibration process, the fit of the common item parameters to the data for each pre-defined group was evaluated. Then, item-by-group interactions were identified when the fit to the data was found to be poor (i.e., the value of the item fit statistic, discussed below, was higher than a chosen threshold value). In the subsequent runs, new *unique or group-specific* item parameters were estimated in the group or groups in which misfit was found and the item fit threshold was gradually lowered until the ultimate target threshold was reached, thus allowing additional group-specific item parameters to be estimated. The fundamental consideration of using this stepwise procedure is to optimize both the model data fit and the comparability across all groups—keeping common item parameters for as many groups as possible or minimizing the use of unique parameters. By allowing unique item parameters for items that show item-by-group interactions – in contrast to excluding such items or accepting poor common item parameter fit – the measurement error is reduced without introducing bias. The research base for this approach can be found in Meredith (1993[14]); Reise, Widaman and Pugh (1993[15]); Glas and Verhelst (1995[16]); Yamamoto (1997[17]); Glas and Jehangir (2014[18]); Meredith and Teresi (2006[19]); as well as Oliveri and von Davier (2011[20]; 2014[21]).

Since PISA 2015, in 2018 and now in 2022, the same IRT calibration and scaling approach has been used to estimate new item parameters onto the existing IRT scales. However, the historical data no longer needed to be included in the scaling since all trend items (reused from 2015 and/or prior PISA cycles) had already been calibrated and scaled. Therefore, in PISA 2022, as in PISA 2018, a fixed item parameter linking approach was utilized with the trend item parameters fixed to their values established in the 2015 and 2018 scaling in the first calibration run to start the estimation of international parameters for the new items. The subsequent runs, then proceeded in the same manner as described above to evaluate item-by-country-by-language interactions (i.e., group-level item-fit) and to estimate unique parameters when needed.

Group-level item-fit analyses are a critical part of the scaling analyses described above. Different types of differential item functioning (DIF) statistics can be used to evaluate the extent to which the IRT model applied to a group fits the response data collected from that group. In the context of the IRT models used in since PISA 2015, the extent to which the model-based item characteristic curve (ICC, computed using formula 11.1 or 11.2 for the 2PLM or the GPCM) and the empirical ICC can differ is evaluated based on the mean deviation (MD) and the root mean square deviation (RMSD) statistics:

Formula 11.5

$$MD_g = \int \left[p_g^{obs}(\theta) - p_g^{exp}(\theta) \right] f_g(\theta) d\theta,$$

Formula 11.6

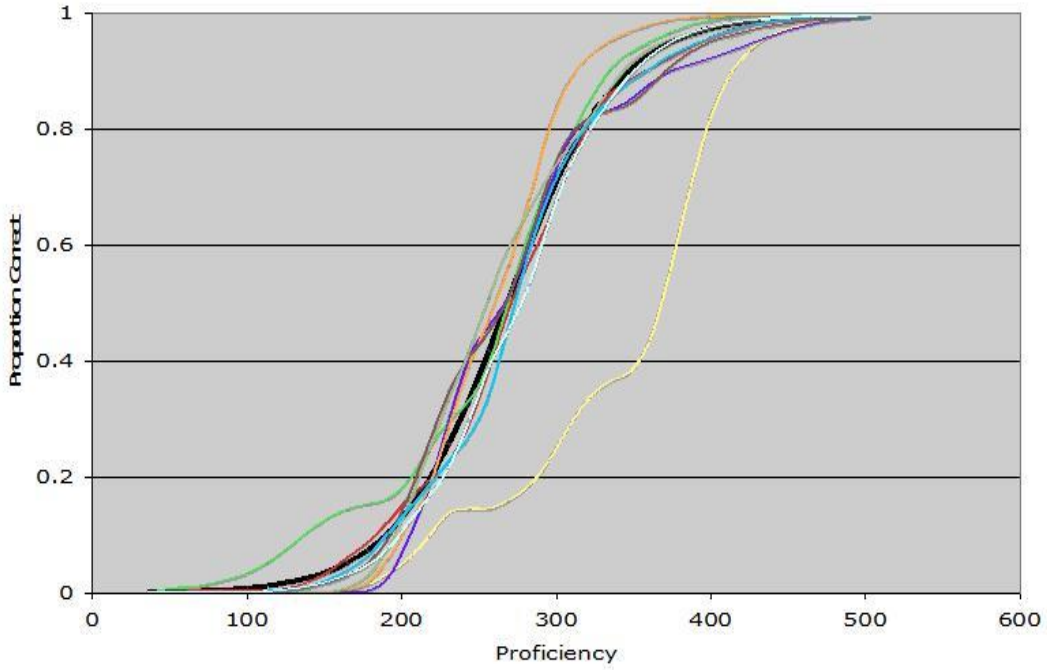
$$RMSD_g = \sqrt{\int \left[p_g^{obs}(\theta) - p_g^{exp}(\theta)\right]^2 f_g(\theta) d\theta},$$

where g = 1, …, G is a country-by-language group; $p_g^{obs}(\theta)$ and $p_g^{exp}(\theta)$ are the observed and expected probability of a correct response given proficiency θ; and $f_g(\theta)$ is the group-specific density on the students' ability scale (Khorramdel, Shin and von Davier, 2019[22]; von Davier, 2005[23]). The observed probability correct is based on the pseudo counts from the expectation-maximization (EM) algorithm that is used to estimate the model (Bock and Aitkin, 1981[24]), while the expected probability correct is based on the estimated item parameters. The moments of the group-specific densities are also estimated for each country-by-language group (Xu and von Davier, 2008[25]).

The observed item characteristic curve (ICC) is obtained from the observed responses across students for each item, and the expected ICCs are computed based on the IRT model using the estimated item parameters. RMSD quantifies the magnitude and MD quantifies the magnitude and direction of deviations in the observed data from the estimated common or group-specific item characteristic curves for each single item. However, while MD is sensitive to the difference in observed and model-based item difficulty represented by the *b* parameter in formulae 11.1 and 11.2, RMSD is sensitive to the differences in both item difficulty and item discrimination represented by the *a* (or slope) parameter in formulae 11.1 and 11.2.

To demonstrate the use of item fit statistics (RMSD, MD), Figure 11.10 shows one example plot for a dichotomously scored item estimated via the 2PLM. It illustrates how the common item parameter fits data from all groups, except for one group. In the figure, the solid black curve is the model-based 2PLM item response curve that corresponds to the common item parameters; the other lines are observed proportions of correct responses along the proficiency scale (horizontal axis) for the data from each group. This plot indicates that the IRT model-based curve conforms to the observed data; proportions of correct responses given the proficiency are quite similar for most countries/economies. However, the data for one country/economy, indicated by the yellow line, shows a noticeable departure from the common item characteristic curve and curves for other groups. This item is far more difficult in that particular country/economy, conditional on proficiency level. Thus, a unique set of parameters would be estimated for this item, for this group.

Figure 11.10. Item response curve (ICC) for an item where the common item parameter is not appropriate for one group



Calibration and scaling of the mathematics and reading adaptive domains

The purpose of adaptive testing is to better match test difficulty with student proficiency and avoid administering items that are either too easy or too difficult. Unlike data collected using traditional linear testing, this results in some of the data (responses to some of the relatively easy or difficult items) being missing not at random and a reduced overlap between test forms delivered to students having different proficiency levels. Unfortunately, using such data for IRT scaling could lead to bias in the item parameters and the student proficiency estimates (Jewsbury and van Rijn, 2020[26]). To address this issue, many testing programs use a two-step data collection design that allows for item parameters to be pre-calibrated through a non-adaptive data collection. Then, once their item parameters have been established, they are incorporated into the operational instrument administration (Glas, 2010[27]). However, for PISA, such approach would require the collection of much larger, population representative, field trial data.

Instead, the PISA reading and mathematics MSATs were designed to ensure both adaptation for many countries/economies performing across wide proficiency ranges, and appropriate data collection for the accurate scaling and estimation of international and unique parameters for all countries/economies. To do so, three issues that could threaten the quality of the reading and mathematics PISA scaling were addressed.

First, in designing and finalizing the MSAT, units were assigned to ensure the linkage across different MSAT forms (i.e., routing paths) through common units appearing multiple times across testlets. Similar to the BIB designs used in earlier PISA cycles, in which the same cluster appears across different forms, such linkage through common units across different testlets was expected to improve the efficiency of the item calibration. Such design considerations were tested and verified with simulation studies before the main survey implementation. Second, a proportion of students were assigned in a non-adaptive manner by overwriting some routing decisions as part of the reading MSAT design or by developing a non-adaptive MSAT assigned to a proportion of students as part of the mathematics *hybrid* MSAT design. In both cases, this ensured that more than 250 responses across the full proficiency range were collected for all items in all countries/economies. Third, the order of position of units within testlets has to vary to be able to adapt

and assemble easier and more difficult testlets. See Chapter 2 for more detailed descriptions of the design implemented.

The effectiveness of the PISA MSAT designs was investigated during their development using data simulation and field trial data, and the quality of the designs implemented was confirmed using main survey data.

Within-testlet unit order effects were examined in the 2022 mathematics and the 2018 reading field trials to confirm the invariance of item parameters by unit order (Yamamoto et al., forthcoming[28]). If the unit order had shown to significantly impact item parameter and proficiency estimates, an MSAT design could not have been implemented because a significant lack of invariance would undermine the effectiveness of the design. The field trial results confirmed the feasibility of introducing an MSAT into the main survey, as unit order effects were found to be negligible.

Model data fit from the same calibration approach used for other non-adaptive domains and alternatives that incorporated MSAT-specific information, such as routing outcomes to define the group in the multi-group calibration process, were evaluated through simulation studies (van Rijn and Shin, 2019[29]). Results showed that incorporating MSAT-specific information in the group definition for the multiple-group IRT model resulted in larger errors in the item parameter estimation. Because routing decisions in PISA are largely based on cognitive responses (i.e., sum scores based on the machine-scored items), using this information again to define groups for the multiple-group IRT model would violate the conditional independence assumptions. In the end, after reviewing the results from calibrating simulated data and the collected main survey data, it was determined that the same approach used for the calibration of the other non-adaptive domains was appropriate. A recent study (Jewsbury et al., 2023[30]) also provides theoretical justification for this choice.

Calibration and scaling of reading fluency

As discussed in Chapter 3, reading fluency items were included as a part of the reading scale, which was assessed principally through the reading MSAT. These items were introduced in 2018 to increase the measurement precision at lower levels of the reading scale. However, as their content and format tend to differ from that of the “regular” reading items, the reading fluency items could affect the existing reading scale. Therefore, following the procedure established in 2018 data ((OECD, 2022[31]), Chapters 9 and 12), to maintain the existing reading scale and avoid any potential issues that could weaken the comparability of the reading scale across cycles, the calibration of reading fluency items was done after the estimation of reading items had been finalized. That is, after the scaling of “regular” reading items was finalized, the reading fluency data was added to the reading data and the reading fluency items were scaled. Because all items were trend, their parameters were fixed to their final 2018 values.

Population modelling and multiple imputation

This section describes the population modelling approach that is employed in the analyses of PISA data that combines the latent regression model for a large number of background variables with the IRT model for cognitive item responses. It also explains the imputation methodology for obtaining plausible values for proficiency (both scales and subscales) and for using these to estimate descriptive statistics for populations and subpopulations. This methodology provides countries/economies with databases that can be used for secondary analyses of relationships between proficiency and background variables.

The prime goal of PISA is to compare the skills and knowledge of 15-year-old students across countries/economies and over cycles, reporting on group-level scores in the core domains of mathematics, reading, and science, as well as other domains (Kirsch et al., 2013[32]). For group-level reporting assessments such as PISA, where the number of items that can be administered to each student is limited

and where the focus of the assessment is on population characteristics, the use of point estimates could lead to seriously biased estimates of population characteristics (Mislevy, 1991[33]; Thomas, 2002[34]; von Davier, Gonzalez and Mislevy, 2009[35]; von Davier et al., 2006[36]; Wingersky, Kaplan and Beaton, 1987[37]).[5] Reporting outcomes are not intended to have consequences of any sort for individual students, and test forms are kept relatively short to minimise the testing burden on students. At the same time, PISA aims to provide a broad content coverage of each of the domains through a large number of items organised into different, but linked, test forms. Thus, each student receives a relatively small number of items from two domains in a two-hour testing period.

Population modelling for PISA 2022 followed the same general approach used in previous cycles. This approach incorporates the IRT scaling of the students' cognitive data from multiple domains, and the students' background data specified as covariates (e.g., gender, country/economy of birth, academic and non-academic activities, attitudes, etc.) through multivariate latent regression models (von Davier et al., 2006[36]). Data from multiple cognitive domains are modelled together to increase the accuracy of the population estimates in each domain by borrowing information from the other cognitive domains. The *plausible value methodology* uses the latent regression models estimated from each country/economy data to impute multiple proficiency values (plausible values) for each student instead of a single point estimate in each domain. The imputation draws the plausible values from the posterior distributions constructed through the multivariate latent regression model and the student data. The multiple imputations from the posterior distributions can then be used to appropriately account for measurement errors in the relations between (sub)population proficiency distributions and characteristics in the background data.

IRT scaling, latent regression, and multiple imputation are carried out through the following steps:

1. *IRT scaling:* estimates the item parameters for each domain to provide comparable scales across countries/economies and cycles using the unidimensional IRT models described in Formula 11.1 and Formula 11.2 (see also section "IRT calibration and scaling").
2. *Latent regression:* estimates the regression coefficients (Γ) and the residual variance-covariance matrix (Σ) using the estimated item parameters from step 1 as true values (Thomas, 1993[38]).
3. *Multiple imputation:* draws ten plausible values for each student on each domain from posterior distributions of proficiency using estimated Γ and Σ (Mislevy and Sheehan, 1987[39]; von Davier, Gonzalez and Mislevy, 2009[35]).

Because of the large number of background collected, a "divide-and-conquer" approach (Patz and Junker, 1999[40]) is used to reduce the computational burden of Step 2 (latent regression) and to avoid over-parametrisation. First, all variables in the BQ are contrast coded.[6] Contrast coding allows for the inclusion of missing responses and avoids the necessity of assuming a linear relationship between the responses to any question and the outcome variable. Second, a principal components analysis (PCA) is conducted to 1) remove collinearity among variables when present and 2) reduce the large number of contrast-coded BQ variables into a smaller number of principal components that are sufficient to account for a large proportion of the variation in the BQ variables without over-parameterisation. This process is conducted country/economy by country/economy to accommodate common BQ variables collected across all countries/economies, to accommodate optional specific BQ variables of participating country/economy's interest, and to allow for the estimation of country/economy-specific relationships between the BQ data and the proficiency variables.

The country/economy-specific multivariate latent regression gives an expression for student's proficiency distributions on the multidimensional scales conditional on covariates (**y**) in addition to the item responses (**x**). Based on Bayes' theorem, the posterior distribution of skills given the observed item responses and covariates (i.e., contextual information) is constructed as follows:

Formula 11.7

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

where $\boldsymbol{\theta}_v$ is a vector of length *D* with scale values (these values correspond to performance on each of the skills) for student *v*. As shown, the posterior distribution of proficiency is proportional to the likelihoods of the item-response data and prior distributions. Given the conditional independence assumption, $P(\boldsymbol{x}_v|\boldsymbol{\theta}_v)$ is the product of independent likelihoods for the observed response to each cognitive item (estimated by IRT models) within each scale (i.e., the likelihood is factored). Next, $P(\boldsymbol{\theta}_v|\boldsymbol{y}_v,\Gamma,\Sigma)$, which is a prior distribution, is the multivariate joint density of proficiencies of the scales, conditional on the extracted principal components derived from background responses, and parameters Γ and Σ. Note that Formula 11.7 technically also depends on the item parameters, but these are treated as fixed in the computations in steps 2 and 3 and therefore dropped from the equation.

More precisely, the latent proficiency variables for each student *v* are assumed to follow multivariate normal distributions:

Formula 11.8

$$\boldsymbol{\theta}_v \sim N_D(\Gamma'\boldsymbol{y}_v, \Sigma),$$

where Γ is the $K \times D$ matrix of regression coefficients, *K* is the number of conditioning variables (the number of principal components plus a dummy for the intercept), and Σ is the $D \times D$ residual variance-covariance matrix. As noted, the parameters Γ and Σ are estimated using the estimated item parameters from the first step. Let $\phi(\theta_v|\Gamma' y_v,\Sigma)$ denote the multivariate normal density with mean $\Gamma' y_v$ and covariance matrix Σ.

Operationally, the procedure is repeated several times to model the main and financial literacy datasets from each country/economy. Once focusing on the core domain data (mathematics, reading, and science; then $D = 4$). Twice focusing on each of the two sets of 4 mathematics subscales data with the reading and science data ($D = 6$). Once focusing on the creative thinking data with the core domains data ($D = 5$). And once focusing on financial literacy with mathematics and reading data ($D = 3$). Latent correlations among those domains are estimated as part of the $D \times D$ residual variance-covariance matrix.

Involving all students in the country/economy, the weighted likelihood function becomes

Formula 11.9

$$L(\Gamma,\Sigma;\ \boldsymbol{X},\boldsymbol{Y}) = \prod_{v=1}^{N} w_v \int \prod_{d=1}^{D} P(\boldsymbol{x}_{vd}|\theta_d)\,\phi(\boldsymbol{\theta}|\Gamma'\boldsymbol{y}_v,\Sigma)d\boldsymbol{\theta},$$

where $\boldsymbol{x}_{vd}$ is the vector of item responses of students for dimension *d*. As noted above, the item parameters $\boldsymbol{\beta}_d$ associated with $P(\boldsymbol{x}_{vd}|\theta_d)$ for dimensions *d=1,…,D* are estimated in the IRT item calibration stage, prior to the estimation of the latent regression $\phi(\boldsymbol{\theta}|\Gamma'\boldsymbol{y}_v,\Sigma)$, and treated as fixed. That is, the latent regression parameters Γ and Σ are estimated conditionally on the previously estimated item parameters $\boldsymbol{\beta}$.

As suggested by Mislevy et al. (1992[41]), the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977[42]) is used for maximizing the likelihood function in Formula 11.9 with respect to Γ and Σ. A multivariate variant of the latent regression model based on the Laplace approximation (Thomas, 1993[38]) is applied in reporting PISA proficiencies on more than two dimensions (domains and subdomains).

After the estimation of regression parameters through the EM algorithm is completed, multiple imputations (plausible values) for each student *v* are drawn from a normal approximation of the conditional posterior distribution of proficiency. More specifically, plausible values are drawn following a three-step process. First, a value for Γ is drawn from $N_D(\hat{\Gamma}, \widehat{V(\Gamma)})$ where $\widehat{V(\Gamma)}$ is the estimated variance of the maximum likelihood estimate $\hat{\Gamma}$ obtained from the EM algorithm (Rubin, 1987[43]). Second, conditional on the generated value for Γ and the fixed value of $\Sigma = \hat{\Sigma}$ obtained from the EM algorithm, the Laplace approximations to the individual posterior mean and variance are computed denoted by $\tilde{\boldsymbol{\theta}}_v$ and $\tilde{\Sigma}_v$, respectively. In the third step, the $\boldsymbol{\theta}_v$ are drawn independently from a multivariate normal distribution $N(\tilde{\boldsymbol{\theta}}_v, \tilde{\Sigma}_v)$ for each student *v* (Chang and Stout, 1993[44]). These three steps are repeated 10 times, effectively resulting in 10 plausible values for $\boldsymbol{\theta}_v$ for each student.

Analysis of data with plausible values

If the multivariate latent proficiencies $\boldsymbol{\theta}_v$ were known for all students, it would be possible to directly compute any statistic $t(\boldsymbol{\theta}, \boldsymbol{y})$, for example, subpopulation sample means, sample percentiles, or sample regression coefficients, to estimate a corresponding population quantity *T*. However, $\boldsymbol{\theta}$ values are not observed, but estimated latent variables through measurement models. To overcome this problem, the approach developed by Rubin (1987[43]) is taken in which $\boldsymbol{\theta}$ is treated as missing data.

Therefore, the value $t(\boldsymbol{\theta}, \boldsymbol{y})$ is approximated by its expectation given the observed data, $(\boldsymbol{x}, \boldsymbol{y})$, as follows:

Formula 11.10

$$t^*(\boldsymbol{x}, \boldsymbol{y}) = E[t(\boldsymbol{\theta}, \boldsymbol{y})|\boldsymbol{x}, \boldsymbol{y}] = \int t(\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{\theta}.$$

It is possible to approximate t^* using plausible values (also referred to as multiple imputations) instead of the unobserved $\boldsymbol{\theta}$ values. A replication approach [see, e.g., Johnson, (1989[45]); Johnson and Rust (1992[46]); Rust, (2014[47])] is used to obtain a variance estimate for the proficiency means of each country/economy and other statistics of interest, and to estimate the sampling variability as well as the imputation variance associated with the plausible values.

As described in the earlier section, plausible values are random draws from the posterior distribution of the proficiencies given the item responses $\boldsymbol{x}_v$, background variables $\boldsymbol{y}_v$, and estimated model parameters. For any student, the value of $\boldsymbol{\theta}_v$ used in the computation of *t* is replaced by a randomly selected value from the student's posterior distribution. Rubin (1987[43]) argued that this process should be repeated several times so that the uncertainty associated with imputation can be quantified. For example, the average of multiple estimates of *t*, each computed from a different set of plausible values, is a numerical approximation of t^* in the above Formula (11.10); the variance among them reflects uncertainty due to not observing $\boldsymbol{\theta}_v$. It should be noted that this variance does not include any variability due to sampling from the population.

It cannot be emphasized strongly enough that the plausible values are not a substitute for individual point estimates (e.g., single test scores). Plausible values are used to make accurate group-level inferences, but they should not be used to make any inferences about individuals. Plausible values are only intermediary computations in the calculation of the expectations in order to estimate population characteristics such as subgroup means and standard deviations. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the individual proficiencies with whom they are associated (Marsman et al., 2016[48]; von Davier, Gonzalez and Mislevy, 2009[35]). Unlike the plausible values, the more familiar ability estimates of educational measurement are optimal for each student (e.g., bias-

corrected maximum likelihood estimates, which are consistent estimates of a student's proficiency, or Bayesian posterior mean estimates, which provide minimum mean-squared errors with respect to a reference population). Point estimates that are optimal for individual students have distributions that can produce decidedly non-optimal and biased estimates of population characteristics (Little and Rubin, 1983[49]). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects. For a further discussion of plausible values, see Mislevy et al. (1992[41]).

Once the plausible values for each students have been produced (in PISA U=10 plausible values are produced for each student for each domain except Creative Thinking, for which 10 plausible scores are generated [7]), they can be employed to estimate the value of a population, subpopulation or group estimator T (e.g., mathematics proficiency) and the magnitude of the errors associated with the estimate as follows:

1. Use the vector made up of the of first of the students' plausible values, and calculate the group estimator T as if the plausible values were the true values of θ. Denote the result T_1.
2. Calculate the sampling variance of T_1. Denote the result $V(T_1)$.
3. Carry out steps 1 and 2 for each of the U vectors of plausible values, thus obtaining T_u and $V(T_u)$ for $u = 1,2,\ldots,U$.
4. The best estimate of the group quantity T is then the average of T_u, obtainable from the U sets of plausible values:

Formula 11.11

$$T. = \frac{\sum_{u=1}^{U} T_u}{U}.$$

1. An estimate of the error variance of the estimator T is the sum of two components, which are the variance due to sampling of examinees and the variance due to latency of the proficiency θ (often called measurement error):

Formula 11.12

$$V(T.) = \frac{\sum_{u=1}^{U} V(T_u)}{U} + \left(1 + \frac{1}{U}\right)\frac{\sum_{u=1}^{U}(T_u - T.)^2}{U-1}.$$

The first component in $V(T.)$ reflects uncertainty due to sampling from the population because PISA samples only a portion of the entire population of 15-year-old students. The second component reflects uncertainty due to measurement error because the students' proficiencies θ are estimated from a limited number of item responses for each respondent.

Example for partitioning the estimated error variance

The following example illustrates the use of plausible values for partitioning the error variance in one country/economy. Table 11.12 presents data for six subgroups of students differing in the context questionnaire variable "Books at home" (variable ST013Q01TA, where 1 = 0-10 books; 2 = 11-25 books; 3 = 26-100 books; 4 = 101-200 books; 5 = 201-500 books; 6 = more than 500 books). Ten plausible values were calculated for each student in a domain. This table presents the means T_{ug} and the sampling standard errors $V(T_{ug})^{1/2}$ for each plausible value (u=1, …,10) and each subgroup defined by the variable ST013Q01TA (g=1,…,6). The bottom section of the table shows the resulting estimates and errors for each subgroup.

Because the standard error associated with the group estimator *T* is comprised of sampling error and measurement error, it can be reduced by either increasing the precision of the measurement instrument or reducing the sampling error. In PISA, a resampling method is used to estimate the sampling variance $V(T_{ug})$, which uses a balanced repeated replication (BRR) approach (See Chapter 10 for details). This component of variance is similar across the ten plausible values; its values are influenced by the homogeneity of proficiencies among students in the subgroup. Note that the sampling error is generally much larger than the measurement error.

Application to the PISA 2022 Main Survey

This section describes the implementation of IRT scaling and population modelling of the PISA 2022 main survey data. Details of the data and procedure implemented, in particular for the mathematics and reading domains that implemented MSAT as well as for the reading fluency items are described first. The dimensionality analyses conducted to verify the applicability of the unidimensional 2PLM and GPCM models to the mathematics MSAT and the innovative creative thinking domains are described next. Then, the country/economy-specific population modelling analyses and the generation of plausible values are detailed. Finally, the procedure utilised to estimate the linking errors between the 2022 and prior PISA cycles is explained.

IRT scaling

IRT scaling is the first step in the modelling of PISA data. It was conducted through the multi-group IRT calibration and scaling approach described earlier, using the international 2022 main survey data and using the trend item parameters fixed to their values established in the previous PISA cycle (common international or unique country-by-language) to ensure appropriate linking to the PISA scale. Each domain was calibrated separately using the *mdltm* software (Khorramdel, Shin and von Davier, 2019[22]; von Davier, 2005[23]) setup to fix already established item parameters and to estimate new ones with the unidimensional 2PLM and GPCM models.

The mathematics and financial literacy assessments included both trend and new items. Reading and science included only trend items. As the innovative domain, creative thinking included only new items. All the PBA and new PBA assessments of mathematics, reading, and science included only trend items, with PBA being the same instruments since 2015 and new PBA being the same instrument as the PISA for Development 2018 instrument (sharing many items in common with PBA) (OECD, 2019[50]).

Table 11.13 details the number of trend and new items kept in the analyses after some items were dropped due to content and/or psychometric reasons that could not be resolved (1 in mathematics, 1 in reading, 1 in financial literacy and 6 items in creative thinking).

The total numbers of students for each domain-specific IRT calibration are detailed in Table 11.14. Calibrations were conducted using the final student weights provided by sampling for each country/economy (Chapter 6, this report) adjusted so the total student weight for each country/economy was 5,000. In this way all participating country/economy contributed equally to the estimation of the new items' international parameters. However, the unweighted number of item responses was used to check whether the minimum number of 250 responses required for evaluation item-by-country-by-language interactions (item-fit) was reached. This was done to ensure that the MD and RMSD statistics could be accurately estimated and the decision to estimate unique parameters when item-misfit was detected appropriate. Nonresponses prior to a student's last valid item response in a cluster were considered omitted and treated as incorrect responses; whereas nonresponses at the end of the cluster were considered not-reached and treated as missing. For CBA mathematics and reading, because of their

MSAT design, the treatment of omit and not-reached responses was done considering the whole test rather than by cluster.

Estimation of common international and group-specific item parameters

Different language versions of the assessment used in countries/economies could result in some items functioning differently in some country-by-language groups. Thus, different language versions of the assessment within a country/economy were treated as separate groups when estimating item parameters. In total, 116 country-by-language groups were used in PISA 2022 multiple-group IRT calibrations for CBA reading, mathematics, and science. In creative thinking and financial literacy 102 and 31 country-by-language groups were analysed, respectively. For PBA and New PBA, 4 countries, each using 1 language were analysed.

To account for cultural and language differences, the stepwise calibration process described earlier was implemented to scale the 2022 data. In the first calibration and fit analyses run, for the trend items, common and group-specific item parameter estimates obtained from the PISA 2018 scaling were used as fixed values. For the new items, common item parameters to all the groups were estimated. Given these parameter estimates, RMSD and MD fit statistics were then computed for all items in all groups, and cases with RMSD above a threshold[8] were identified.

In the relatively rare instances where large RMSD misfit was found (values above 0.4), the item was dropped in the specific group (i.e., excluded from scaling in that group). In the subsequent calibrations and fit analyses runs, unique parameters were estimated, as long as there were 250 unweighted responses, gradually lowering the RMSD threshold to 0.12—a value that was found to be optimal for maximizing both the overall model-data fit and the proportion of international item parameters across country-by-language groups (Joo et al., 2019[51]). A review of the results obtained in the final calibration run was also conducted to identify any case where even with unique parameters estimated a value below RMSD of 0.18 could not be reached or very low slope parameter (below 0.1) or extreme difficulty parameters (above 5 in absolute value) were obtained. When such cases were found, the item was dropped in the specific group or specific groups.

In addition to ensuring appropriate model fit and reducing the measurement error, maintaining the comparability of scales through common item parameters across countries/economies, assessment modes, and assessments over time is of prime importance. Therefore, the *mdltm* software used for item calibration implements an algorithm that monitors RMSD and MD across the specified groups and suggests a list of items to be re-estimated for each group. This algorithm seeks to minimize the number of group-specific item parameters needed to fit the data. It does so, item by item, constraining the item parameters to be the same across the groups in which the item exhibits misfit in the same direction (positive or negative). Thus, the same specific item parameters may be unique to one group or multiple groups (e.g., country-by-language groups) exhibiting similar misfit patterns. Ultimately, through the iterative process it may be discovered that the unique parameters common to more than one group need to be relaxed further and re-estimated separately to reach the desired fit. But this is done only when needed so that the total number of unique parameters is minimized across all countries/economies.

Dimensionality analyses

The results of the scaling analyses just described show that the IRT models used, with the unidimensionality and local independence assumptions, do fit the data quite well. However, it was important to further evaluate these assumptions for the major and the innovative domains, which included a large proportion of newly developed items and all newly developed items.

Residual analyses of field trial mathematics and creative thinking data and residual analysis of main survey creative thinking data were conducted for each country/economy to assess both the conditional

independence and unidimensionality assumptions. For mathematics, additional dimensionality analyses of the main survey data were conducted to verify that the new items developed based on the revised framework do not introduce a new dimension, distinct from the one captured by the mathematics PISA scale developed in prior PISA cycles. This was done by fitting a two-dimensional IRT simple-structure model which treated trend and new items as two different latent traits and evaluating the extent to which the more complex two-dimensional model of the total weighted data from all countries/economies provided a significant improvement in fit. These analyses were conducted in the same way as in previous cycles for the major domain of mathematics and the innovative creative thinking domain (OECD, 2017[3]; 2020[52]). The methods implemented to conduct residual analysis are detailed below; results are reported in the next sections.

The *mdltm* software (von Davier, 2005[23]) computes residuals in the step that follows the item calibration. For dichotomous item responses, response residuals for a person *v* with estimated ability $\hat{\theta}_v$ for each item *i = 1,..., n* were defined as below:

Formula 11.13

$$r(x_{vi}) = \frac{x_{vi} - P(X_i = 1 \mid \hat{\theta}_v)}{\sqrt{P(X_i = 1 \mid \hat{\theta}_v)\left[1 - P(X_i = 1 \mid \hat{\theta}_v)\right]}}.$$

For polytomous item responses, response residuals were calculated using the conditional mean and variance defined below:

Formula 11.14

$$r(x_{vi}) = \frac{x_{vi} - E(X_i \mid \hat{\theta}_v)}{\sqrt{V(X_i \mid \hat{\theta}_v)}},$$

Formula 11.15

$$E(X_i \mid \hat{\theta}_v) = \sum_{k=1}^{m_i} kP(x_{vi} = k|\hat{\theta}_v),$$

Formula 11.16

$$V(X_i \mid \hat{\theta}_v) = \sum_{k=1}^{m_i} k^2 P(x_{vi} = k|\hat{\theta}_v) - \left[E(X_i \mid \hat{\theta}_v)\right]^2.$$

Once the item response residuals have been calculated, the item residual correlations across respondents can be computed to produce an item residual correlation matrix. Although the null distribution of such residual correlations--also known as the Q_3 statistic (Yen, 1984[53]) — are not well known, unidimensional and locally independent data are expected to show random residual correlations patterns around zero across all items and across items within each unit (Chen and Thissen, 1997[54]; Yen, 1984[53]). Local item dependencies are found when an item pair shows highly correlated response residuals and their item slope parameter estimates are high. In such cases where an item pair or multiple item pairs within a unit show

local item dependence, this may be addressed by scoring these two items or the whole unit as a single polytomous score and modelled with the partial or generalized partial credit model described earlier in this chapter (Rosenbaum, 1988[55]; Wilson and Adams, 1995[56]).

Following the inspection of the residual correlation matrix and the treatment of local item dependences, principal component analysis of the residual correlation matrix was conducted to evaluate the extent to which the instrument is unidimensional. If the unidimensionality assumption holds, little common variance among the item response residuals is expected after the ability dimension has been accounted for by the IRT model. In this case, a principal component analysis will produce a scree plot where no single component accounts for much more variance than any other.

Mathematics dimensionality analyses

Residual-based dimensionality analyses of the CBA mathematics were conducted on the field trial data to identify potential local item dependence and to confirm the unidimensionality of the mathematics instrument assembled for the main survey. Based on the item-by-item correlations for all mathematics items, no item pairs were identified with exceptionally strong correlations. Furthermore, the unidimensional IRT scaling analyses of the field trial data and later the main survey data (as described above) did not show any items with unusually large slope parameters. Both IRT scaling and residual analysis provided evidence that the conditional independence assumption was not violated.

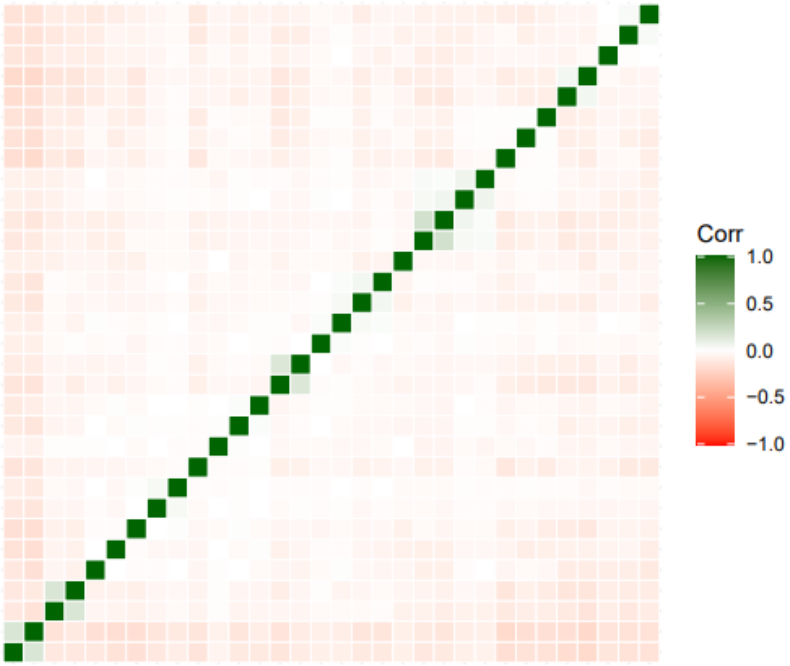
The two-dimensional IRT modelling of the mathematics main survey data, where trend and new items were assigned to two different latent proficiency scales, provided an additional check of the unidimensionality assumption. When the multidimensional IRT model was fitted, the trend item parameters were fixed to the common international item parameters obtained from the PISA 2018 cycle, and the new items were constrained to the newly estimated unidimensional international parameters. Although the Akaike Information Criterium (AIC) (Akaike, 1974[57]) showed better fit for the two-dimensional model, the Bayesian Information Criterium (BIC) (Schwarz, 1978[58]) and the log-penalty improvement showed that the unidimensional model fits better and the multidimensional model provides very little improvement over the unidimensional model (Table 11.15). In particular, it was found that the unidimensional model reached 99.8% of the model fit improvement over the independence model compared to the gains expected from the multidimensional model. Similarly, the two-dimensional IRT model of the field trial data showed only marginal improvement in overall model fit over the unidimensional IRT model. Moreover, the correlations of two sets of group means (the trend item only and the new items only) from the multidimensional model were very high, ranging from 0.91 to 0.99 across the different country-by-language groups. Additionally, the dimension-specific weighted likelihood estimates (WLEs) of student ability were very highly correlated with the unidimensional WLEs.

Considering all the evidence gathered from the field trial and main survey data analyses, there is strong evidence that the new and trend mathematics items and scores can be placed on the existing unidimensional PISA scale.

Creative thinking dimensionality analyses

As the innovative domain, creative thinking was an entirely new domain in 2022. Field trial analyses showed that the instrument was essentially unidimensional. For the main survey, 36 items were selected out of the 40-item field trial item pool. The unidimensional IRT scaling of the main survey data was conducted and response residuals were calculated. Pairwise residual item correlations were then computed for each country-by-language group and averaged across groups. Figure 11.11 shows the residual correlation matrix obtained. Besides the dark green squares on the diagonal that represent each item correlating with itself, no strong pairwise residual correlation and no noticeable patterns that could be indicative of additional dimension(s) was observed.

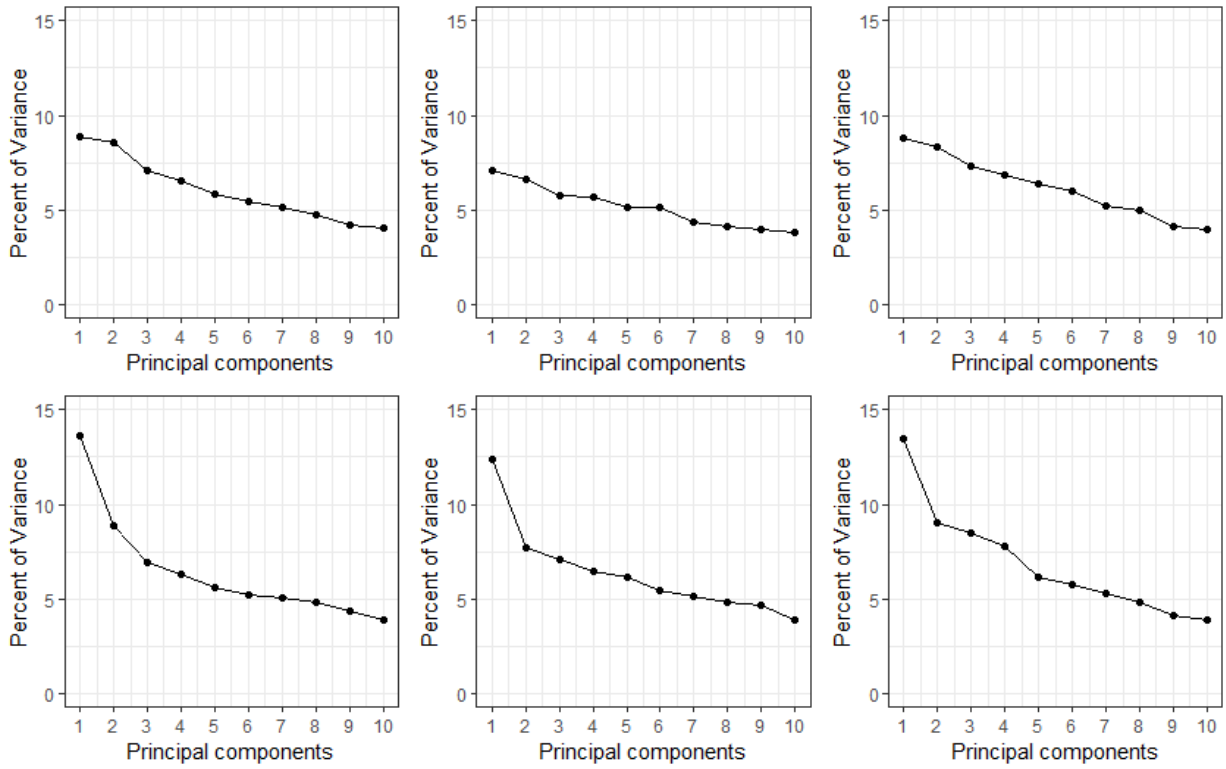
Figure 11.11. Residual correlation matrix for the creative thinking main survey



As part of the residual analysis, principal components of the residual correlation matrix were extracted. Should the eigenvalue of the first principal component be much larger than the other principal components, an additional latent trait, other than the overall ability, could be present. When all the item residual correlations are included as variables, the percentage of variance adds up to 100%. Analysis results across countries/economies, showed that the percentage of variance for the first principal component ranges from 7.1% to 13.7% with a median of 10.2% and the percentage of variance accounted for by the first 10 principal components ranges from 50.8% to 73.65%, with a median value of 63.14%. Thus, the first component did not account for a large part of the variance accounted for by the first ten components. This was confirmed by inspection of each country/economy principal component analysis scree plots in most cases.

The plots in Figure 11.12 show six countries' scree plots as the most distinctive examples. In most cases illustrated by the top three scree plots no clear "elbow" that would be indicative of an additional dimension not accounted for by the unidimensional IRT scaling. However, in a few cases some evidence of multidimensionality was observed. Nevertheless, overall, the results supported the scaling and reporting of creative thinking as proficiency using a unidimensional scale.

Figure 11.12. Percentage of variance from principal component analyses for 6 countries/economies



Population modelling in PISA 2022

The population model described earlier was applied to the PISA 2022 data. Fixing the item parameters to their values obtained from the unidimensional IRT scaling, multivariate latent regression models were fitted to the data at the country/economy level, and 10 plausible values per domain were generated for each student. Plausible values for core domains (reading, mathematics, and science) were generated for all students participating in the assessment, regardless of whether they were administered items in that domain. Plausible values for the innovative domain were generated for all students if countries/economies opted for the CrT domain. That is, students received plausible values for each test domain administered in their country/economy according to the test design implemented regardless of the specific forms they took. Students who did not participate or did not have responses in a particular domain were assigned model-dependent plausible values for that domain based on their responses to the BQ as well as the cognitive responses in other domains.

Measurement errors must be considered when dealing with the plausible values in the secondary analyses. The plausible values for the domain(s) students did not take have larger uncertainty than the plausible values for the other domains that were administered to them. By using repeated analysis with each of the 10 plausible values, the measurement error will readily be reflected in the analyses and the final aggregation of results can be conducted in a way that the variability across the 10 analyses is properly reflected.

While most covariates used in the population modelling come from the student BQ responses, some additional covariates were derived from the cognitive assessment's process data. Same as done in PISA 2018 (see Annex H of the PISA 2018 technical report), such derived covariates include response time information, and school-level WLEs to capture the unique variations across schools, which are relevant for predicting proficiency distributions within each country/economy.

The following sections provide further information about how the population model was applied to PISA 2022 data, how plausible values were generated, and how plausible values can be used in further analyses.

Main sample, creative thinking and financial literacy sample models

The software called DGROUP (Educational Testing Service, 2012[59]) was used to estimate the multivariate latent regression models and generate plausible values (von Davier and Sinharay, 2014[12]; von Davier et al., 2006[36]). During the estimation, the item parameters for the cognitive items were fixed at the values obtained from the multi-group IRT models described earlier in this chapter. As in previous PISA cycles, nearly all student BQ variables, as well as some contextual characteristics, were included.

All BQ variables were contrast-coded before they were processed further. The contrast coding scheme is reproduced in Annex B of this report. Contrast coding allows for the inclusion of missing responses and avoids the necessity of assuming a linear relationship between the responses to any question and the outcome variable. Note that with the introduction of within-construct matrix sampling design, missing by matrix-sampling design and missing by omitting behavior were distinguished, which increased the number of contrast codes for BQ variables. With contrast-coded BQ variables, a PCA is conducted to 1) remove collinearity among variables when present and 2) reduce the large number of contrast-coded BQ variables into a smaller number of principal components that are sufficient to account for a large proportion of the variation in the BQ variables without over-parameterisation. Because each country/economy can have unique associations among the BQ variables, a set of principal components was calculated for each country/economy. As such, the extraction of principal components was carried out separately by country/economy. In PISA, the number of principal components retained in each of the multivariate latent regression models was selected to be the smaller of 1) the number of principal components needed to explain 80% of the BQ variance, and 2) the number that corresponds to 6.7% (1/15) of the raw sample size. Note that in previous PISA 2015 and 2018 cycles, the number that corresponds to 5% (1/20) of the raw sample size was used. However, with the increase in BQ scales and variables, the rule was relaxed to retain more information in the extracted principal components. Still, this avoided a numerical instability in the estimation that could occur due to potential overparameterization of the model.

The main sample data collection included the core domains administered by all 81 participating countries/economies and the innovative creative thinking domain administered by 64 countries/economies. Separate population modelling analyses of the core domains, of mathematics subscales with reading and science, and of creative thinking with the core domains were conducted. The financial literacy sample data collection was offered as an international option and was administered by 20 countries/economies. The cognitive instruments included trend items from 2012, 2015, and 2018, and a few new items. For the population modelling, the financial literacy sample (who took Forms 67 – 74) was combined with the students from the main sample who took reading and mathematics only (Forms 1 – 12). This was done to establish a stable linkage between the financial literacy and main PISA forms, and the reading and mathematics domains. Thus, the financial literacy sample received plausible values in mathematics, reading, and financial literacy, but not in science and not in mathematics subscales.

Treatment of students with fewer than six test item responses

This section addresses the issue of students who provided background information but did not respond to enough cognitive items. Students with responses to fewer than six cognitive items in any domain were not included in the multivariate latent regression modelling to avoid unstable estimations of the Γ and Σ.

In PISA 2022, fewer than: 0.09% of students were excluded from the core domains CBA or new PBA multivariate latent regressions; 7.4% the mathematics sub-scales; 0.04% the creative thinking; and less than 0.03% from the financial literacy multivariate latent regressions. Nevertheless, the population model

was applied to these students for the generation of plausible values. For each of the two mathematics subscales (by *process* and by *content*), the proportion of students excluded from the modelling is larger because responses to at least six items in the relatively short subscales were needed to be included in the multivariate latent regression model.

Consistent with the data treatment applied in the IRT scaling, nonresponses prior to a valid response were considered omitted and treated as incorrect responses; whereas nonresponses at the end of each of the cluster (for non-adaptive domains) or each MSAT session (for mathematics and reading) were considered not-reached and treated as missing in the population modelling and PV generation.

Plausible values

Plausible values for the domains evaluated were drawn from the normal approximations to the posterior distributions estimated from the multivariate latent regression models.

The plausible value variables for the domains follow the naming convention PV1<domain> through PV10<domain>, where "<domain>" took on the following form:

- MATH for mathematics
- READ for reading
- SCIE for science
- CRTH_NC[9] for creative thinking
- FLIT for financial literacy

Population modelling for the mathematics subscales

The aim of generating plausible values for the different mathematics subscales is to provide proficiency estimates representative of important aspects within the overall mathematics framework. These subscales allow for secondary analyses of relationships between proficiency and BQ variables that focus of different aspects within the mathematics domain. However, it should be noted that subscales proficiencies (plausible values) are based on fewer items than the full scale and, thus, are associated with larger measurement error.

There were two sets of subscales reported for mathematics. These were process subscales related to mathematical reasoning (employing mathematical concepts, facts, and procedures; interpreting, applying, and evaluating mathematical outcomes; formulating situations mathematically; reasoning) and content subscales related to mathematical content knowledge (space and shape; quantity; change and relationships; uncertainty and data). Mathematics subscales were computed for the CBA only. Table 11.16 gives an overview of the 233 (one item was dropped) mathematics items by the cognitive process and the test structure. It should be noted that the two mathematics subscale category types are based on a two-way classification of the same 233 items (distributed into the 4 + 4 = 8 subscales). In other words, each item contributed to one of the cognitive process subscales and one of the content subscales.

Because the cognitive process subscales and the content subscales were based on the same set of mathematics items, population modelling for the cognitive process subscales and the population modelling for the content subscales could only be done separately. Therefore, two additional multidimensional population models were fitted for each CBA country/economy to provide the desired mathematics subscale PVs. These two models were:

- Model 1: reading, science, and the four subscales of mathematics cognitive process, thus, 6 dimensions in total;

- Model 2: reading, science, and the four subscales of mathematics content subscales, thus, 6 dimensions in total.

Reading and science data were used for the population modelling of the mathematics subscales to maximize the information used from the students. PVs were generated for those domains (reading and science) in these runs, but only the PVs for the mathematics subscales were included in the database for each set of mathematics subscales.

The item parameters used for the population modelling of the mathematics subscales were the same as those for the overall mathematics scale described above, which were obtained from the unidimensional multi-group IRT model for mathematics. Therefore, the mathematics subscales and the overall mathematics scale proficiencies can be compared as they are on the same scale. However, because the mathematics scale is not the weighted average of the mathematics subscales, a country/economy's mean proficiency in mathematics can be noticeably different from the country/economy's mean subscale proficiencies.

The plausible values reported for the mathematics subscales follow the naming convention PV1<subscale> through PV10<subscale>, where "<subscale>" takes on the following form:

- MCCR Content Subscale of Mathematics – Change and Relationships
- MCQN Content Subscale of Mathematics – Quantity
- MCSS Content Subscale of Mathematics – Space and Shape
- MCUD Content Subscale of Mathematics – Uncertainty and Data
- MPEM Cognitive Process Subscale of Mathematics – Employing Mathematical Concepts, Facts, and Procedures
- MPFS Cognitive Process Subscale of Mathematics – Formulating Situations Mathematically
- MPIN Cognitive Process Subscale of Mathematics – Interpreting, Applying, and Evaluating Mathematical Outcomes
- MPRE Cognitive Process Subscale of Mathematics – Reasoning

Finally, as noted earlier, PVs from the same draw should be used when assessing correlations between domains or when conducting secondary analyses, not from different draws. Thus, estimating correlations between MPEM1, MPFS1, MPIN1, MPRE1 is appropriate, while estimating correlations between MPEM1, MPFS2, MPIN3, MPRE4 is inappropriate. The same is true for the content subscale. Because the core domain PVs and the subscale PVs reported were draws from different population models, estimating correlations between them would not be appropriate. However, the correlations between the other cognitive domains and the subscales that are part of the each one of the two subscale population models estimated are reported in Chapter 14.

Linking PISA 2022 to previous PISA cycles

There are three measurable sources of error variance to account for when using the PISA data. These are error due to student sampling, error due to the reliability of the assessment, and error due to the linking of different instruments across assessment cycles.

Following the approach implemented in 2015, an evaluation of the magnitude of linking error was conducted by considering differences between reported country/economy results from previous PISA cycles and the transformed results from rescaling prior to 2015. The magnitude of the linking errors is related to the changing assessment framework, instruments, mode of delivery and scaling methods over PISA cycles. It is also related to changes from major to minor domain that could leads to a recombination of items and units within clusters, as well as to changes in design from linear to adaptive.

As in past cycles, scale-level differences across countries/economies between adjacent calibrations are considered as the target of inference. The effect of the variability of two calibrations is evaluated at the cross-country/economy level, while within-country/economy sampling variability is not targeted. Moreover, sampling variance and measurement variance are two separate variance components that are accounted for by the variance estimation based on replicate weights and plausible values. Taken together, the focus of the linking error lies on the expected variability on the country/economy mean over the different calibrations.

The definition of calibration differences starts from the ability estimates of a respondent v from country/economy g in a target cycle under two separate calibrations (e.g., the original calibration of a PISA cycle and its recalibration), C1 and C2. We can write for calibration C1:

Formula 11.17

$$\tilde{\theta}_{v,C1,g} = \theta_{v,true} + \hat{u}_{C1,g} + \tilde{e}_v,$$

where $\hat{u}_{C1,g}$ denotes the estimated country/economy specific error term in C1 and $\tilde{e}_v$ is the respondent specific measurement error; and for calibration C2 accordingly:

Formula 11.18

$$\tilde{\theta}_{v,C2,g} = \theta_{v,true} + \hat{u}_{C2,g} + \tilde{e}_v.$$

Defined in this way, there may be country/economy level differences in the expected values of respondents based on the calibration. These are a source of uncertainty and can be viewed as adding variance to country/economy-level estimates. Given the assumption of a country/economy-level variability of estimates due to C1 and C2 calibrations, for the differences between estimates we find:

Formula 11.19

$$\tilde{\theta}_{v,C1,g} - \tilde{\theta}_{v,C2,g} = \hat{u}_{C1,g} - \hat{u}_{C2,g},$$

and the expectation can be estimated by:

Formula 11.20

$$E\left(\hat{u}_{C1,g} - \hat{u}_{C2,g}\right) = \tilde{\mu}_{g,C1} - \tilde{\mu}_{g,C2} = \hat{\Delta}_{C1,C2,g}.$$

Across countries/economies, the expected differences of country/economy means ($\tilde{\mu}$) can be assumed to vanish, since the scales are transformed after calibrations to match distribution moments. That is, we may assume:

Formula 11.21

$$\sum_{g=1}^{G} E\left(\hat{u}_{C1,g} - \hat{u}_{C2,g}\right) = 0 = \sum_{g=1}^{G} \hat{\Delta}_{C1,C2,g}.$$

The variance of the differences of country/economy means based on C1 and C2 calibrations can then be considered the linking error of the trend comparing the Y2 cycle means that were used to obtain calibration C2 estimates, and the Y1 cycle estimates. The linking error can be written as:

Formula 11.22

$$V[\hat{\Delta}_{C1,C2,g}] = \frac{1}{G}\sum_{g=1}^{G}(\tilde{\mu}_{g,C1} - \tilde{\mu}_{g,C2})^2.$$

The main characteristics of this approach can be summarised as follows:

- Scale-level differences across countries/economies from adjacent-cycle IRT calibrations C1 and C2 are considered.
- The effect of the variability of scale-level statistics between two calibrations is evaluated at the country/economy level.
- Within-country/economy sampling variability is not targeted.
- Sampling variance and measurement error are two separate variance components that are accounted for by plausible values and replicate weights-based variance estimation.

The use of this variance component is analogous to that of previous cycle linking errors. The variance calculated in the formula (11.22) is a measure of uncertainty due to re-estimation of the model when using additional data from subsequent cycles, obtained with potentially different assessment designs, estimation methods, and underlying databases. To avoid the possibility that some data points (countries/economies) have excessive influence on the results, the robust S_n statistic was used, as it was in PISA 2015 and 2018. The S_n statistic was proposed by Rousseeuw and Croux (1993[2]) as a more efficient alternative to the scaled median absolute deviation from the median (1.4826*MAD) that is commonly used as a robust estimator of standard deviation. It is defined as:

Formula 11.23

$$S_n = 1.1926 * med_i\left(med_j(|x_i - x_j|)\right).$$

The differences defined above are plugged into the formula, that is, $x_{i=\hat{\Delta}_{C1,C2,i}}$ are used to calculate the linking error for comparisons of cycles Y1 and Y2 based on calibrations C1 (using only Y1 data) and C2 (using Y2 data and additional data including Y1). The robust estimates of linking error between cycles by domain are presented in Chapter 14.

The S_n statistic is available in SAS as well as the R package "robustbase." See also *https://cran.r-project.org/web/packages/robustbase/robustbase.pdf*.

References

Akaike, H. (1974), "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, Vol. 19/6, pp. 716–723, https://doi.org/10.1109/TAC.1974.1100705. [57]

Beaton, A. (ed.) (1987), *Joint estimation procedures*, Educational Testing Service. [37]

Birnbaum, A. (1968), “Some latent trait models and their use in inferring an examinee’s ability”, in Lord, F. and M. Novick (eds.), *Statistical Theories of Mental Test Scores*, Addison-Wesley. [6]

Bock, R. and M. Aitkin (1981), “Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm”, *Psychometrika*, Vol. 46/4, pp. 443-459. [24]

Bock, R. and M. Zimowski (1997), “Multiple group IRT”, in van der Linden, W. and R. Hambleton (eds.), *Handbook of Modern Item Response Theory*, Springer-Verlag. [4]

Chang, H. and W. Stout (1993), “The asymptotic posterior normality of the latent trait in an IRT model”, *Psychometrika*, Vol. 58/1, pp. 37 - 52, https://doi.org/10.1007/BF02294469. [44]

Chen, W. and D. Thissen (1997), “Local dependence indexes for item pairs using item response theory”, *Journal of Educational and Behavioral Statistics*, Vol. 22/3, pp. 265–289, https://doi.org/10.3102/10769986022003265. [54]

Dempster, A., N. Laird and D. Rubin (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm”, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39/1, pp. 1-38, https://www.jstor.org/stable/2984875. [42]

Educational Testing Service (2012), *DGROUP [Computer software]*. [59]

Fischer, G. and I. Molenaar (eds.) (1995), *Rasch Models: Foundations, Recent Developments, and Applications*, Springer. [9]

Glas, C. (2010), “Item Parameter Estimation and Item Fit Analysis”, in van der Linden, W. and C. Glas (eds.), *Elements of Adaptive Testing*, Springer. [27]

Glas, C. and K. Jehangir (2014), “Modelling Country Specific Differential Item Functioning”, in Rutkowski, L., M. von Davier and D. Rutkowski (eds.), *Handbook of International Large-scale Assessment*, CRC Press. [18]

Glas, C. and N. Verhelst (1995), “Testing the Rasch Model”, in Fischer, G. and I. Molenaar (eds.), *Rasch Models: Foundations, Recent Developments, and Applications*, Springer. [16]

Jewsbury, P. et al. (2023), *Modeling multistage and targeted testing data with item response theory*, [Manuscript submitted for publication], Research and Development Division, Educational Testing Service. [30]

Jewsbury, P. and P. van Rijn (2020), “IRT and MIRT models for item parameter estimation with multidimensional multistage tests”, *Journal of Educational and Behavioral Statistics*, Vol. 45/4, pp. 383-402. [26]

Johnson, E. (1989), “Considerations and techniques for the analysis of NAEP data”, *Journal of Educational Statistics*, Vol. 14/4, pp. 303-334, https://doi.org/10.3102/10769986014004303. [45]

Johnson, E. and K. Rust (1992), “Population inferences and variance estimation for NAEP data”, *Journal of Educational Statistics*, Vol. 17/2, pp. 175–190, https://doi.org/10.3102/10769986017002175. [46]

Joo, S. et al. (2019), *Evaluating Item Fit Statistic Thresholds in PISA: The Analysis of Cross-Country Comparability of Cognitive Items*, [Manuscript submitted for publication], Research and Development Division, Educational Testing Service. [51]

Khorramdel, L., H. Shin and M. von Davier (2019), "GDM software mdltm including parallel EM algorithm", in von Davier, M. and Y. Lee (eds.), *Handbook of Psychometric Models for Cognitive Diagnosis*, Springer. [22]

Kirsch, I. et al. (2013), "On the growing importance of international large-scale assessments", in von Davier, M., E. Gonzalez and I. Kirsch (eds.), *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*, Springer, https://doi.org/10.1007/978-94-007-4629-9_1. [32]

Leys, C. et al. (2013), "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median", *Journal of Experimental Social Psychology*, Vol. 49/4, pp. 764–766, https://doi.org/10.1016/j.jesp.2013.03.013. [1]

Little, R. and D. Rubin (1983), "On jointly estimating parameters and missing data", *American Statistician*, Vol. 37/3, pp. 218–220. [49]

Marsman, M. et al. (2016), "What can we learn from plausible values?", *Psychometrika*, Vol. 81/2, pp. 274–289, https://doi.org/10.1007/s11336-016-9497-x. [48]

Meredith, W. (1993), "Measurement invariance, factor analysis and factorial invariance", *Psychometrika*, Vol. 68/4, pp. 525–543, https://doi.org/10.1007/BF02294825. [14]

Meredith, W. and J. Teresi (2006), "An essay on measurement and factorial invariance", *Medical Care*, Vol. 44/11, pp. S69–S77, https://doi.org/10.1097/01.mlr.0000245438.73837.89. [19]

Mislevy, R. (1991), "Randomization-based inference about latent variables from complex samples", *Psychometrika*, Vol. 56/2, pp. 177–196, https://doi.org/10.1007/BF02294457. [33]

Mislevy, R. et al. (1992), "Estimating population characteristics from sparse matrix samples of item responses.", *Journal of Educational Measurement*, Vol. 29/2, pp. 133–161, https://doi.org/10.1111/j.1745-3984.1992.tb00371.x. [41]

Mislevy, R. and K. Sheehan (1987), "Marginal Estimation Procedures", in Beaton, A. (ed.), *Implementing the New Design: The NAEP 1983-84 Technical Report*, Educational Testing Service. [39]

Muraki, E. (1992), "A generalized partial credit model: Application of an EM algorithm", *Applied Psychological Measurement*, Vol. 16/2, pp. 159-177, https://doi.org/10.1002/j.2333-8504.1992.tb01436.x. [7]

OECD (2022), *PISA 2018 Technical Report*. [31]

OECD (2020), *PISA 2018 Technical Report*, PISA, OECD Publishing, Paris, https://www.oecd.org/pisa/data/pisa2018technicalreport/. [52]

OECD (2019), *PISA for Development Technical Report*, OECD Publishing, Paris, http://www.oecd.org/pisa/pisa-for-development/pisafordevelopment2018technicalreport/. [50]

OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris, http://www.oecd.org/pisa/data/2015-technical-report/. [3]

Oliveri, M. 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, https://doi.org/10.1080/15305058.2013.825265. [21]

Oliveri, M. and M. von Davier (2011), “Investigation of model fit and score scale comparability in international assessments”, *Psychological Test and Assessment Modelling*, Vol. 53/3, pp. 315–333. [20]

Patz, R. and B. Junker (1999), “A straightforward approach to Markov chain Monte Carlo methods for item response models”, *Journal of Educational and Behavioral Statistics*, Vol. 24/2, pp. 146 - 178, https://doi.org/10.2307/1165199. [40]

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Nielsen and Lydiche. [8]

Reise, S., K. Widaman and R. Pugh (1993), “Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance”, *Psychological Bulletin*, Vol. 114/3, pp. 552–566, https://doi.org/10.1037/0033-2909.114.3.552. [15]

Rosenbaum, P. (1988), “Permutation tests for matched pairs with adjustments for covariates.”, *Applied Statistics*, Vol. 37/3, pp. 401–411, https://doi.org/10.2307/2347314. [55]

Rousseeuw, P. and C. Croux (1993), “Alternatives to the median absolute deviation”, *Journal of the American Statistical Association*, Vol. 88/424, pp. 1273–1283, https://doi.org/10.2307/2291267. [2]

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons. [43]

Rust, K. (2014), “Sampling, weighting, and variance estimation in international large-scale assessments”, in Rutkowski, L., M. von Davier and D. Rutkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, CRC Press. [47]

Rutkowski, L., M. von Davier and D. Rutkowski (eds.) (2014), *Analytics in International Large-Scale Assessments: Item Response Theory and Population Models*, CRC Press. [12]

Schwarz, G. (1978), “Estimating the Dimension of a Model”, *Annals of Statistics*, Vol. 6/2, pp. 461 - 464. [58]

Thomas, N. (2002), “The role of secondary covariates when estimating latent trait population distributions”, *Psychometrika*, Vol. 67/1, pp. 33–48, https://doi.org/10.1007/BF02294708. [34]

Thomas, N. (1993), “Asymptotic corrections for multivariate posterior moments with factored likelihood functions”, *Journal of Computational and Graphical Statistics*, Vol. 2/3, pp. 309–322. [38]

van der Linden, W. and R. Hambleton (eds.) (2016), *Handbook of Modern Item Response Theory*, Springer. [11]

van der Linden, W. and R. Hambleton (1997), “Item Response Theory: Brief History, Common Models, and Extensions”, in van der Linden, W. and R. Hambleton (eds.), *Handbook of Modern Item Response Theory*, Springer. [10]

van Rijn, P. and H. Shin (2019), *Item Calibration for Multistage Tests in the Context of Large-Scale Educational Assessment*, [Manuscript in preparation], Research and Development Division, Educational Testing Service. [29]

von Davier, M. (2005), "A general diagnostic model applied to language testing data", *Research Report No. RR-05-16*, Educational Testing Service. [23]

von Davier, M., E. Gonzalez and R. Mislevy (2009), "What are plausible values and why are they useful?", *IERI Monograph Series*, No. 2/1, http://www.ierinstitute.org/IERI_Monograph_Volume_02_Chapter_01.pdf. [35]

von Davier, M. et al. (2006), "Statistical Procedures Used in the National Assessment of Educational Progress (NAEP): Recent Developments and Future Directions", in Rao, C. and S. Sinharay (eds.), *Handbook of Statistics: Psychometrics*, Elsevier, https://doi.org/10.1016/S0169-7161(06)26032-2. [36]

von Davier, M. and K. Yamamoto (2004), "Partially observed mixtures of IRT models: An extension of the generalized partial credit model", *Applied Psychological Measurement*, Vol. 28/6, pp. 389–406, https://doi.org/10.1177/0146621604268734. [5]

von Davier, M. et al. (2019), "Evaluating item response theory linking and model fit for data from PISA 2000-2012", *Assessment in Education: Principles, Policy & Practice*, Vol. 26/4, pp. 466-488, https://doi.org/10.1080/0969594X.2019.1586642. [13]

Wilson, M. and R. Adams (1995), "Rasch models for item bundles", *Psychometrika*, Vol. 60/2, pp. 181–198, https://doi.org/10.1007/BF02301412. [56]

Xu, X. and M. von Davier (2008), "Comparing multiple-group multinomial log-linear models for multidimensional skill distributions in the general diagnostic model", *ETS Research Report Series*, No. ETS RR-08-35, Educational Testing Service, Princeton, NJ. [25]

Yamamoto, K. (1997), "Scaling and scale linking", in Murray, T., I. Kirsch and L. Jenkins (eds.), *Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey*, National Center for Education Statistics. [17]

Yamamoto, K. et al. (forthcoming), "Improved test designs and multistage adaptive testing in large-scale assessments", in Khorramdel, L. et al. (eds.), *Innovative Computer-Based International Large-Scale Assessments: Foundations, Methodologies and Quality Assurance Procedures*, Springer. [28]

Yen, W. (1984), "Effects of local item dependence on the fit and equating performance of the three-parameter logistic model", *Applied Psychological Measurement*, Vol. 8/2, pp. 125–145, https://doi.org/10.1177/014662168400800201. [53]

Notes

1. More detail on the parallel MSAT Reading and Mathematics can be found in Chapter 2 of this Technical Report.

2. With MSAT, testlets of different difficulty are assembled specifically for each stage (core, stage 1 and stage 2), therefore position effects cannot easily be compared across stages.

3. Computed using senate weights so that all countries/economies contribute equally.

4. Note that the parameterisations $(\theta_v - b_i + d_{ir})$ and $(\theta_v - b_{ir})$, both used in the IRT literature, are equivalent. However, the former has the advantage of using b_i with both the 2PLM and GPCM, representing the overal item difficulty.

5. In contrast, tests that are used to report individual-level results are concerned with accurately assessing the performance of each individual test-taker for the purposes of diagnosis, selection, or placement. This is achieved by administering a relatively large number of items to each individual, resulting in a negligible level of uncertainties associated with the point estimates.

6. The contrast variables derived from the BQ responses can be found in the Annex B to this Technical Report

7. As the mathematical properties of both plausible values and scores (the latter being obtained via a non-linear transformation of the former), plausible values will be used throughout the chapter for brevity.

8. Note that RMSD are always larger than absolute MD values. Therefore, unless one wishes to set different thresholds on RMSD and MD to identify misfit, it is sufficient to use a single threshold on RMSD.

9. Population modeling and plausible values are first produced on each domain's IRT theta scale and then transformed to each domain's reported PISA scale. All domains other than creative thinking use a linear transformation. Creative thinking uses a non-linear test characteristic curve transformation that results in plausible values that correspond to the student's plausible number correct (NC) on a form made up of all the items in the creative thinking item pool.

Chapter 11 tables

Table 11.1. Language(s) of assessment, mode of assessment, and number of students and schools sampled for each country/economy

Country	Language(s)	Test Mode	Main Sample	Financial Literacy Sample	Total	Schools
Albania (ALB)	Albanian	CBA	6,156		6,156	283
Argentina (ARG)	Spanish	CBA	12,127		12,127	460
Australia (AUS)	English	CBA	13,521		13,521	761
Austria (AUT)	German	CBA	6,159	1,599	7,758	304
Baku (Azerbaijan) (QAZ)	Azeri, Russian	CBA	7,720		7,720	199
Belgium* (BEL)	French, German, Dutch	CBA	8,286	1,189	9,475	285
Brazil (BRA)	Portuguese	CBA	10,810	2,901	13,711	602
Brunei Darussalam (BRN)	English	CBA	5,576		5,576	54
Bulgaria (BGR)	Bulgarian	CBA	6,118	1,605	7,723	203
Cambodia (KHM)	Khmer	New PBA	5,279		5,279	183
Canada* (CAN)	French, English	CBA	23,386	4,203	27,589	885
Chile (CHL)	Spanish	CBA	6,489		6,489	231
Chinese Taipei (TAP)	Chinese	CBA	5,896		5,896	188
Colombia (COL)	Spanish	CBA	7,804		7,804	262
Costa Rica (CRI)	Spanish	CBA	6,122	1,453	7,575	199
Croatia (HRV)	Croatian	CBA	6,135		6,135	180
Cyprus (QCY)	Greek, English	CBA	6,517		6,517	102
Czech Republic (CZE)	Czech	CBA	8,460	2,213	10,673	430
Denmark (DNK)	Danish, Faroese	CBA	6,224	1,578	7,802	349
Dominican Republic (DOM)	Spanish	CBA	6,902		6,902	254
El Salvador (SLV)	Spanish	CBA	6,705		6,705	290
Estonia (EST)	Russian, Estonian	CBA	6,392		6,392	196
Finland (FIN)	Finnish, Swedish	CBA	10,256		10,256	242
France (FRA)	French	CBA	6,771		6,771	283
Georgia (GEO)	Georgian, Azerbaijani, Russian	CBA	6,583		6,583	267
Germany (DEU)	German	CBA	7,712		7,712	259
Greece (GRC)	Greek	CBA	6,545		6,545	235
Guatemala (GTM)	Spanish	New PBA	5,190		5,190	290
Hong Kong (China) (HKG)	Chinese, English	CBA	6,048		6,048	168
Hungary (HUN)	Hungarian	CBA	6,236	1,639	7,875	263
Iceland (ISL)	Icelandic	CBA	3,367		3,367	136
Indonesia (IDN)	Indonesian	CBA	13,471		13,471	412
Ireland (IRL)	Irish, English	CBA	5,569		5,569	170
Israel (ISR)	Hebrew, Arabic	CBA	6,251		6,251	193
Italy (ITA)	Italian, German	CBA	10,564	2,789	13,353	345
Jamaica (JAM)	English	CBA	3,956		3,956	154
Japan (JPN)	Japanese	CBA	5,760		5,760	182
Jordan (JOR)	Arabic	CBA	7,799		7,799	260
Kazakhstan (KAZ)	Kazakh, Russian	CBA	19,768		19,768	571
Korea (KOR)	Korean	CBA	6,454		6,454	186
Kosovo (KSV)	Serbian, Albanian	CBA	6,027		6,027	229
Latvia (LVA)	Latvian, Russian	CBA	5,394		5,394	226
Lithuania (LTU)	Lithuanian, Russian, Polish	CBA	7,257		7,257	292
Macao (China) (MAC)	English, Chinese, Portuguese	CBA	4,384		4,384	46
Malaysia (MYS)	Malay, English	CBA	7,069	1,818	8,887	199
Malta (MLT)	Maltese, English	CBA	3,127		3,127	46
Mexico (MEX)	Spanish	CBA	6,288		6,288	280
Mongolia (MNG)	Kazakh, Mongolian	CBA	6,999		6,999	195
Montenegro (MNE)	Montenegrin, Albanian	CBA	5,800		5,800	64
Morocco (MAR)	French, Arabic	CBA	6,867		6,867	178
Netherlands (NLD)	Dutch	CBA	5,046	1,278	6,324	154

Country	Language(s)	Test Mode	Main Sample	Financial Literacy Sample	Total	Schools
New Zealand (NZL)	English	CBA	4,830		4,830	175
North Macedonia (MKD)	Macedonian, Albanian	CBA	6,610		6,610	111
Norway (NOR)	Nynorsk, Bokmål	CBA	6,616	1,719	8,335	266
Palestinian Authority (PSE)	Arabic, English	CBA	7,905		7,905	273
Panama (PAN)	Spanish, English	CBA	4,590		4,590	227
Paraguay (PRY)	Spanish	New PBA	5,087		5,087	283
Peru (PER)	Spanish	CBA	6,968	1,819	8,787	336
Philippines (PHL)	English	CBA	7,193		7,193	188
Poland (POL)	Polish	CBA	6,048	1,574	7,622	246
Portugal (PRT)	Portuguese	CBA	6,819	1,805	8,624	226
Qatar (QAT)	Arabic, English	CBA	7,676		7,676	229
Republic of Moldova (MDA)	Russian, Romanian	CBA	6,235		6,235	265
Romania (ROU)	Romanian, Hungarian	CBA	7,364		7,364	262
Saudi Arabia (SAU)	Arabic, English	CBA	6,928	1,829	8,757	193
Serbia (SRB)	Hungarian, Serbian	CBA	6,432		6,432	185
Singapore (SGP)	English	CBA	6,608		6,608	165
Slovak Republic (SVK)	Slovak, Hungarian	CBA	5,833		5,833	289
Slovenia (SVN)	Slovenian	CBA	6,752		6,752	350
Spain (ESP)	Catalan, Galician, Basque, Spanish, Valencian	CBA	30,920	1,682	32,602	983
Sweden (SWE)	Swedish, English	CBA	6,079		6,079	263
Switzerland (CHE)	German, French, Italian	CBA	6,847		6,847	262
Thailand (THA)	Thai	CBA	8,507		8,507	280
Türkiye (TUR)	Turkish	CBA	7,250		7,250	196
Ukrainian regions (QUR)	Ukranian	CBA	4,005		4,005	176
United Arab Emirates (ARE)	Arabic, English	CBA	24,623	6,452	31,075	843
United Kingdom (Excl. Scotland) (QUK)	Welsh, English	CBA	9,932		9,932	345
United Kingdom (Scotland) (QSC)	English	CBA	3,277		3,277	120
United States (USA)	English	CBA	4,602	1,121	5,723	160
Uruguay (URY)	Spanish	CBA	6,747		6,747	230
Uzbekistan (UZB)	Karakalpak, Uzbek, Russian	CBA	7,293		7,293	202
Viet Nam (VNM)	Vietnamese	PBA	6,137		6,137	180

Note: Ukranian regions (QUR) - 18 out of 27 regions administered the assessment.
*Denotes a country/economy for which the financial literacy domain was not fully sampled across the population; it is not a nationally-representative sample.

Table 11.2. Example output for examining response distributions

BLOCK M01 (UNWEIGHTED)
Response Analysis

Which plan best represents the drawing o

		1	NOT RCH	OFF TSK	OMIT	0	1			TOTAL	R BIS =	0.6064
ITEM 1	N	1	14	74	2054	5466				7608	PT BIS =	0.4551
	PERCENT	0.01	0.18	0.97	27.00	71.85				100.00	P+ =	0.7185
CM033Q01S	MEAN SCORE	7.00	5.00	1.22	3.59	7.31				6.25	DELTA =	10.69
	STD. DEV.	0.00	3.09	1.87	2.92	3.49				3.75		
MAC	RESP WT	0.00	0.00	0.00	0.00	1.00					ITEM WT =	1.00

Which is the third fastest time?

		2	NOT RCH	OFF TSK	OMIT	0	1			TOTAL	R BIS =	0.6213
ITEM 2	N	8	0	98	2204	5299				7601	PT BIS =	0.4722
	PERCENT	0.11	0.00	1.29	29.00	69.71				100.00	P+ =	0.6971
CM474Q01S	MEAN SCORE	1.25	0.00	1.38	3.66	7.42				6.25	DELTA =	10.94
	STD. DEV.	1.48	0.00	1.66	3.06	3.14				3.75		
MAC	RESP WT	0.00	0.00	0.00	0.00	1.00					ITEM WT =	1.00

How many people (boys and girls combined

		3	NOT RCH	OFF TSK	OMIT	00	11	12	13	21	TOTAL	R BIS =	0.8431
ITEM 3	N	20	1	1139	1639	335	530	201	3744		7589	PT BIS =	0.7118
	PERCENT	0.26	0.01	15.04	15.01	4.41	6.98	2.65	49.33		100.00	P+ =	0.5636
DM155Q02C	MEAN SCORE	1.00	3.00	2.58	2.58	5.40	5.81	6.33	8.81		6.26	DELTA =	12.36
	STD. DEV.	0.55	0.00	2.02	2.02	2.48	2.69	2.71	2.84		3.74		
HUM	RESP WT	0.00	0.00	0.00	0.00	0.50	0.50	0.50	1.00			ITEM WT =	2.00

Table 11.3. Example table of item score category analysis and item flags summary

BLOCK M01 (UNWEIGHTED)
Item Score Category Analysis (Partial credit model)

	Category	N	Pct. At	Pct. Below	Mean	Std. Dev.	Biserial	B *
ITEM 1	0	2142	28.15	0.00	3.52	2.93		
CM033Q01S	1	5466	71.85	28.15	7.31	3.49	0.6064	-0.9529
ITEM 2	0	2302	30.29	0.00	3.57	3.05		
CM474Q01S	1	5299	69.71	30.29	7.42	3.14	0.6213	-0.8303
ITEM 3	0	2779	36.62	0.00	3.01	2.31		
DM155Q02C	1	1066	14.05	36.62	5.78	2.65	0.6114	0.3033
	2	3744	49.33	50.67	8.81	2.84	0.5728	-0.8367

BLOCK M01 (UNWEIGHTED)
Item Analysis Flag Summary

Item ID	Num Resp	Type	R-BIS	P-PLUS	% NOTRCH	% OFFTSK	% OMIT	% MISS	Flags
CM033Q01	2	SCR	0.6064	0.7185	0.01	0.18	0.97	1.17	.....
CM474Q01	2	SCR	0.6213	0.6971	0.11	0.00	1.29	1.39	.....
DM155Q02	5	ECR	0.8431	0.5636	0.26	0.01	15.01	15.25	...O..

Table 11.4. Flagging criteria for items in the item analyses

	Criteria for flagging items
min rbis/rpoly	0.3
min P+	0.2
max P+	0.9
max Omit%	10
max Offtask%	10
max Not-Reached%	10

Table 11.5. Percentage of response time outliers by domain

DOMAIN	Reading	Science	Mathematics	Financial Literacy	Creative Thinking
Number of Clusters/testlets	30 MSAT testlets	6	144 MSAT testlets	2	5
Number of Outliers	0.68%	1.21%	0.95%	0.53%	0.76%

Note: Statistics for mathematics, reading, science, and creative thinking are based on the main sample; statistics for financial literacy are based on the financial literacy sample.

Table 11.6a. Descriptive statistics for testlet or cluster response time (in minutes)

DOMAIN	N	MIN	Q1	MEDIAN	Q3	MAX	MEAN	SD
Math Testlet 1	543,174	0.04	11.86	16.32	21.31	33.94	16.64	6.80
Math Testlet 2	556,894	0.02	9.86	14.02	17.97	33.94	13.88	5.86
Math Testlet 3	541,703	0.01	6.48	10.17	13.63	33.90	10.16	5.05
Reading Testlet 1	238,303	0.04	9.90	13.64	17.91	31.85	14.04	6.24
Reading Testlet 2	241,238	0.05	13.11	17.67	22.04	37.78	17.44	6.73
Reading Testlet 3	232,806	0.00	6.35	10.44	14.04	37.61	10.29	5.29
Science	236,767	0.03	15.49	21.35	27.47	48.49	21.82	9.50
Financial Literacy	41,682	0.03	15.87	21.90	30.40	53.79	23.22	10.88
Creative Thinking	143,429	0.07	13.60	18.84	24.47	43.21	19.25	8.25

Note: Statistics for mathematics, reading, science, and creative thinking are based on the main sample; statistics for financial literacy are based on the financial literacy sample.

Table 11.6b. Descriptive statistics for domain stage response time (in minutes)

DOMAIN	N	MIN	Q1	MEDIAN	Q3	MAX	MEAN	SD
Mathematics Linear	136,377	0.04	32.23	43.07	50.43	91.06	40.27	12.99
Mathematics MSAT	406,552	0.04	32.52	43.49	50.71	93.85	40.58	12.97
Reading Design A	178,444	0.04	34.91	44.72	50.41	94.60	41.45	12.48
Reading Design B	59,183	0.09	35.02	44.86	50.34	94.06	41.41	12.57
Reading Design A (with RF)	173,578	0.04	34.92	44.76	50.45	94.60	41.46	12.48
Reading Design B (with RF)	57,601	0.09	35.07	44.92	50.38	94.06	41.45	12.57
Science	236,767	0.05	36.17	46.65	53.70	101.34	43.73	12.88
Financial Literacy	41,682	0.11	40.92	49.10	53.32	104.69	45.84	11.57
Creative Thinking	143,429	0.06	30.08	40.35	48.47	93.46	38.52	12.61

Note: Statistics for mathematics, reading, science, and creative thinking are based on the main sample; statistics for financial literacy are based on the financial literacy sample.

Tables 11.7a – 11.7d. Median domain response time (in minutes) by proficiency level

Refer to Chapter_11_Tables_xlsx to view Tables 11.7a to 11.7d online.

Table 11.8a. Median response time (in minutes) by cluster position in the CBA for non-adaptive domains

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Science	29.69	18.08	23.87	17.77	-11.91
Financial Literacy	32.97	16.82	27.38	17.55	-15.42
Creative Thinking	23.81	17.38	20.43	15.98	-7.83

Note: Excludes cluster outliers.

Table 11.8b. Median response time (in minutes) by assessment hour in the CBA for all domains

DOMAIN	1st Hour	2nd Hour	2nd Hour - 1st Hour
Math Linear	45.65	40.03	-5.61
Math MSAT	46.05	40.43	-5.62
Reading Core Items	15.17	12.29	-2.89
Reading Stage 1 and 2	29.44	28.08	-1.37
Reading MSAT	46.75	42.12	-4.64
Science	49.9	42.98	-6.92
Financial Literacy	50.46	46.92	-3.54
Creative Thinking	42.96	37.61	-5.35

Note: Excludes cluster outliers.

Table 11.9a. Average proportion correct (P+) by cluster position in the CBA for non-adaptive domains

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Science	0.452	0.399	0.423	0.384	-0.068
Financial Literacy	0.510	0.434	0.479	0.422	-0.089
Creative Thinking	0.479	0.453	0.456	0.428	-0.051

Table 11.9b. Average proportion correct (P+) by assessment hour in the CBA for all domains

DOMAIN	1st Hour	2nd Hour	2nd Hour - 1st Hour
Math Linear - trend*	0.376	0.357	-0.019
Math Linear - new*	0.388	0.375	-0.013
Math MSAT - trend*	0.376	0.356	-0.02
Math MSAT - new*	0.397	0.385	-0.012
Reading Core Items	0.569	0.524	-0.044
Reading Stage 1 and 2	0.495	0.474	-0.021
Creative Thinking	0.466	0.442	-0.024

Table 11.10a. Average proportion of omitted responses by cluster position in the CBA for non-adaptive domains

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Science	0.027	0.049	0.042	0.06	0.033
Financial Literacy	0.027	0.063	0.041	0.073	0.045
Creative Thinking	0.042	0.041	0.054	0.052	0.009

Table 1111.10b. Average omission rate by assessment hour in the CBA for all domains

DOMAIN	1st Hour	2nd Hour	2nd Hour - 1st Hour
Math Linear - trend*	0.08	0.098	0.017
Math Linerar - new*	0.067	0.074	0.007
Math MSAT - trend*	0.071	0.09	0.019
Math MSAT - new*	0.048	0.058	0.01
Reading Core Items	0.036	0.052	0.015
Reading Stage 1 and 2	0.063	0.078	0.015
Creative Thinking	0.041	0.053	0.012

Table 11.11a. Average proportion correct (P+) by cluster position in new PBA

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Reading	0.623	0.614	0.601	0.583	-0.040
Science	0.480	0.481	0.468	0.462	-0.018
Mathematics	0.387	0.385	0.373	0.356	-0.029

Table 11.11b. Average proportion of omitted responses by cluster position in new PBA

DOMAIN	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
Reading	0.046	0.049	0.055	0.067	0.022
Science	0.061	0.055	0.067	0.074	0.013
Mathematics	0.095	0.090	0.095	0.110	0.015

Table 11.12. Example for use of plausible values for partitioning the error

Plausible value	0-10 books at home		11-25 books at home		26-100 books at home		101-200 books at home		201-500 books at home		500+ books at home	
	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)	Mean	(s.e.)
1	429.16	3.51	473.20	3.19	512.84	2.32	538.82	2.74	559.98	2.93	547.44	4.79
2	429.91	3.38	474.43	3.24	512.68	2.42	539.22	2.63	559.50	3.09	546.99	4.75
3	429.99	3.57	474.13	3.22	513.51	2.40	537.97	2.65	561.92	2.94	546.52	4.44
4	429.34	3.39	475.64	3.35	513.31	2.41	538.97	2.45	559.42	3.01	545.47	4.97
5	429.87	3.42	473.92	3.24	512.92	2.42	539.68	2.54	559.51	3.04	546.58	4.75
6	429.04	3.25	474.58	3.34	513.29	2.43	536.60	2.59	562.07	3.05	546.57	4.66
7	429.35	3.54	474.59	3.35	513.04	2.40	539.21	2.67	559.83	3.05	546.16	4.94
8	429.21	3.41	475.42	3.17	512.85	2.51	541.71	2.60	560.24	3.05	546.25	4.71
9	428.76	3.42	473.17	3.10	512.36	2.36	537.66	2.92	559.86	3.19	547.96	4.64
10	429.50	3.43	473.77	3.04	512.25	2.35	538.45	2.64	560.68	3.04	547.98	4.90

Estimate	429.41	474.29	512.91	538.83	560.30	546.79
Sampling Error	3.43	3.23	2.40	2.65	3.04	4.76
Measurement Error	0.42	0.87	0.43	1.42	1.02	0.85
Standard Error	3.46	3.34	2.44	3.00	3.21	4.83

Table 11.13. Number of trend (linking) items and new items by domain and mode of assessment

	CBA Trend	CBA New	CBA Total	PBA	New PBA
Mathematics	74	159*	233	71	64
Reading	196*		196	87	66
Science	115		115	85	66
Reading Fluency	65		65		79
Financial Literacy	40*	5	45		
Creative Thinking		32*	32		

Note: *Dropped items: CMA112Q02, CR547Q07S, DF082Q01C, and DT520Q01C, DT560Q01C, DT560Q02C, DT450Q01C, DT450Q02C and DT450Q03C

Table 11.14. Unweighted calibration sample size by domain and mode of assessment

	CBA	PBA and New PBA
Mathematics	561,556	15,768
Reading	245,800	13,401
Science	245,715	13,209
Financial Literacy	42,068	
Creative Thinking	144,492	

Table 11.15. Model selection criteria for the unidimensional and the two-dimensional IRT models for trend and new mathematics items in the main survey

MODEL	# of Parameters	AIC	BIC	Log Penalty	Improvement
Independence	NA	NA	NA	0.668	NA
Unidimensional	751	11861255	11869412	0.5794	99.8%
Two-dimensional	1002	11812371	11823248	0.5792	100.0%

Note: Log penalty (Gilula & Haberman, 1994) provides the negative expected log likelihood per observation, the % Improvement compares the log-penalties of the models relative to the difference between most restrictive and most general model.

Table 11.16. Distribution of the items to the mathematics subscales

Content Scale			Process Scale		
Subscales	Trend	New	Subscales	Trend	New
Change and Relationships	17	38	Employing Mathematical Concepts, Facts and Procedures	24	51
Quantity	21	55	Formulating Situations Mathematically	11	37
Space and Shape	17	26	Interpreting, Applying and Evaluating Mathematical Outcomes	10	47
Uncertainty and Data	19	41	Reasoning	29	25
Total:	74	160	Total:	74	160

Note: CMA112Q02S (Content Scale - Quantity; Process Scale - Reasoning) was included in the counts above but was ultimately dropped during scaling for all countries.