Chapter 9
Scaling PISA data

OVERVIEW

The test designs for PISA 2018 and prior cycles were based on an incomplete block design, a type of an incomplete multiple-matrix sampling design in which each student is administered a subset of items from the total item pool. Specifically, in most cycles, a balanced incomplete block (BIB) design was used. With the BIB design, items are grouped into clusters (i.e., sets of items), and these clusters are used to assemble test forms. 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.

A new feature in the test design for reading in PISA 2018 was the use of the multistage adaptive testing (MSAT) design. In this design, the item selection for students was largely based on their performance on the previous items (Yamamoto, Shin, & Khorramdel, 2018, 2019). As a result, instead of administering a set of items to a random group of students, different sets of items were administered to groups of students with different levels of proficiency. That is, students who performed well in the initial stage were administered relatively more difficult items in the subsequent stage, while students who did not perform well in the initial stage were administered easier items. In this situation, statistics based on the number or percent of items correct would need to be adjusted not only for the relative difficulty of the set of items administered to the students, but also for the relative proficiency of the students. Unadjusted item-by-item reporting would ignore the differences in the proficiencies of the subgroups to which the set of items was administered, and thus, would not provide comparable information across populations or subpopulations.

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 matches 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 that accounts for the sources of uncertainty in the data.

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 2018 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 2018 main survey and the previous PISA cycles are explained.

This chapter also describes the approaches used to scale the MSAT data, as well as the results of analyses conducted to evaluate the quality of the new MSAT design, the data collected, and the comparability of the MSAT results with results from prior non-adaptive PISA cycles.

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 in order 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 2018 computer platform had successfully delivered, captured, and exported information for more than 600 computer-based assessment (CBA) and 250 paper-based assessment (PBA) items.

Target sample size, routing, and data yield

Target sample size

The assessment design for the PISA 2018 main survey included the core domains of reading, mathematics, and science, delivered through both CBA and PBA. In addition, it also included the optional domains of financial literacy (FL) and the innovative domain of global competence (GC), delivered only through CBA. As part of the sample 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 GC 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 or the CBA without GC needed to sample 35 students from each of the 150 schools for a total sample of 5,250 students. CBA countries/economies taking the FL domain were also required to sample larger numbers of schools and/or more students per school to obtain an additional sample of 1,650 students. This group of 1,650 students who took FL, referred to as the "FL sample," are randomly equivalent to, albeit different from, the "main sample" students who did not take FL. Note that this was different from the approach used in the 2015 cycle when FL was administered to a subset of the main sample students (see Chapter 2 for more details).

With reading as the major domain, 1 hour of reading was administered to all students in the main sample, and the other domains were only administered to a subset of students.

Data yield

Table 9.1 shows the assessment languages and the sample sizes for each of the participating countries/economies. In order 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., ST012 or ST013).

Table 9.1 Language(s) of assessment, mode of assessment, and number of students and schools sampled for each country/economy

Figures 9.1, 9.2, and 9.3 show the extent to which each country/economy participating in the CBA, the FL 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. A few countries/economies did not reach the sample size requirements because of their small total population size.

Figure 9.1 Main sample yield for countries/economies participating in the CBA

Figure 9.2 Financial literacy sample yield for participating countries/economies

Figure 9.3 Main sample yield for countries/economies participating in the PBA

Since the sample sizes varied greatly from country/economy to country/economy, the number of sampled schools and the sample sizes from each school varied as well. As shown in Table 9.1, the number of schools ranged from 44 (Luxembourg) to 1,089 (Spain), but most countries/economies met the requirement to sample a minimum of 150 schools.

The PISA 2018 assessment design also required that students be randomly assigned to forms in the prescribed proportions. Results showed that this standard was met for all participating countries/economies and that the assignment of students to items was appropriate for the item analyses and IRT scaling.

Reading MSAT data yield

The goal of the reading MSAT design 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 to students as designed and intended. Note that the reading MSAT was not delivered to all students. For example, some students in some countries/economies took a shorter non-adaptive *Une-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.

Three critical aspects of the MSAT design were closely monitored: 1) whether students were randomly assigned to each MSAT core testlet, 2) whether students were randomly assigned to

either Design A (75%) or Design B (25%), and 3) whether students were routed to the Stage 1 and Stage 2 testlets according to the MSAT design (Design A or Design B).

First, results confirmed that the random assignment at the Core stage worked well, with a uniform distribution of about 12.5% of students assigned to each of the eight different testlets. Second, Designs A and B were assigned in the desired proportions (i.e., 75% and 25%, respectively) in each of the participating countries/economies. Third, the more complex routing of students from the Core to Stage 1 and from Stage 1 to Stage 2 was evaluated in detail. Figure 9.4 shows the proportion of students in each country/economy that were routed to the four different testlet combinations: 1) difficult testlets (i.e., High) in both Stage 1 and Stage 2 (HH), 2) a difficult testlet in Stage 1 and an easy testlet (i.e., Low) in Stage 2 (HL), 3) an easy testlet in Stage 1 and a difficult testlet in Stage 2 (LH), and 4) easy testlets in both Stage 1 and Stage 2 (LL). Students were categorized as missing/undetermined when they did not complete certain stages. In the figure below, 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 (sky blue), while in the higher-performing countries/economies, a smaller proportion of students were assigned to the easiest testlets (red). Also, as intended, every type of testlet was assigned to at least 23% of the total sample in each country/economy in each stage, regardless of the proficiency distribution in the country/economy. For the lowest performing countries/economies, adding HH (sky blue) to HL (purple) and HH (sky blue) to LH (green) in Figure 9.4 shows that the lowest percentage of students in these countries/economies were assigned to a difficult testlet in Stage 1 and/or Stage 2. For the highest performing countries/economies, adding LL (red) to LH (green) and LL (red) to HL (purple) shows that the lowest percentage of students in these countries/economics were assigned to an easy testlet in Stage 1 and/or Stage 2. This 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 all range of item difficulties.

Figure 9.4 Proportion of students routed to each testlet combination in Reading MSAT

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
- item difficulty
- item discrimination

Item response categories included several types of non-responses 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 (mathematics, science, FL, and GC) or in the MSAT session for reading. 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 reading form. 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. *Une-heure* (UH) booklet results were provided for countries/economies, where applicable.

Tables 9.2 and 9.3 show examples of the item analysis outputs. Table 9.2 shows the first three items in block/cluster M01. 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 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 9.2 Example output for examining response distributions

Table 9.3 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 9.4. In this example, the third item is flagged for having an omit rate of greater than 10%.

Table 9.3 Example table of item score category analysis and item flags summary

Table 9.4 Flagging criteria for items in the item analyses

Reading MSAT equated P+

In 2018, an *equated P+* statistic was developed and added to the statistics on reading provided to the CBA countries/economies. This equated P+ was necessary because of the MSAT design for reading in which samples of students responding to items were no longer randomly equivalent. Subsamples of students routed to the more difficult MSAT testlets are, on average, more proficient than the total sample of students. The reverse is true for the easier MSAT testlets. Therefore, while classical observed P+ and other statistics are still helpful for identifying items with potential scoring or other problems, observed P+ values are no longer comparable between adaptive and non-adaptive designs (i.e., across cycles within countries/economies), across items of different expected difficulty, and across countries/economies within a cycle.

The equated P+ is equivalent to the P+ that would have been obtained from the non-adaptive designs used in previous PISA cycles. It accounts for the differences in proficiencies between the sample that responded to the item and the total sample. Therefore, the 2018 reading MSAT equated P+ can be compared with the classical observed P+ from 2015 (or earlier cycles) as well as with other equated P+ statistics from different countries/economics that participated in PISA 2018. Computational details for the equated P+ statistic are provided later in this chapter (see the 'IRT modelling and scaling' section).

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[1]. 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, they had to follow the prescribed order of

[1] Note that two item response time variables are reported in the public use files (PUF): one reported using the item ID with the suffix T (T variable), and one using the item ID with the suffix TT (TT variable). Considering a student could come back to revisit an item after visiting other items in the unit, the T variable captures the time spent during the last visit to the item alone, whereas the TT variable captures the aggregate time across all visits to the item. This last variable provides a better accounting of the total time a student may have spent responding to the item. The response time analyses results reported in this chapter are based on the TT variable.

clusters or MSAT stages and units at their own pace. Except for reading (which followed the MSAT design), students were expected to complete two 30-minute clusters within an hour, regardless of the positions within the assessment (e.g., cluster 1 and 2 in the first hour, cluster 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 had to complete the reading fluency items within a 3 minute limit and were expected to complete the reading MSAT items within the remaining time in the hour.

Focusing on larger-than-expected cluster response times, outliers were identified using the median absolute deviation (MAD) approach (Leys, Ley, Klein, Bernard, & Licata, 2013; Rousseeuw & Croux, 1993). 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, 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 9.5 shows the percentages of response time outliers by domain. The proportions of outliers were small—less than 2 percent across all domains. Since reading fluency was short and strictly time-limited, an outlier analysis was not needed for reading fluency.

Table 9.5 Percentage of response time outliers by domain

Cluster (or testlet) level response time

Table 9.6a presents descriptive statistics for the cluster response times for mathematics, science, GC, and FL. These values are the sum of the time each student spent on each item in a cluster, aggregated across students, countries/economies, and positions. The mean and standard deviation of the cluster response times were similar across domains for all countries/economies taking the CBA. On average, students spent about 21 minutes to respond to items in each cluster, with more than 75% of the students completing the cluster in under 29 minutes. With the outliers removed, no student in any country/economy took longer than 60 minutes to finish a given 30-minute cluster. Note that it was possible for students to take close to an hour to complete a given 30-minute cluster and have very little or no time to finish the subsequent cluster with which it was paired. That is, for mathematics and science, when a pair of clusters was administered before or after the mid-test break, the use of up to 60 minutes for the first of the two clusters would leave little to no time to finish the second cluster. Such long response times pointed to potential administration issues. Very short cluster response times of less than 1 minute also pointed to potential administration issues, technical problems with the data collection, or that a student had advanced very rapidly through the items.

Table 9.6.b presents descriptive statistics for the reading fluency and reading MSAT testlet response times. Most students responded to the reading fluency items in less than 1 minute. With the outliers removed, no student took more than 2 minutes. Regarding reading MSAT, students took between 12.3 to 16.5 minutes on average to respond to each of the testlets.[2] Testlet 2 included more items, and as expected, required more time. Nevertheless, most students spent less than 20 minutes on any single testlet. Considering reading fluency and reading MSAT

[2] Note that "Reading Testlet 1" indicates Core, "Reading Testlet 2" indicates Stage 1 for Design A or Stage 2 for Design B, and "Reading Testlet 3" indicates Stage 2 for Design A or Stage 1 for Design B. Details about the MSAT reading design are presented in Chapter 2.

together, students spent less than 43 minutes on average on the reading domain, most students spent less than 50 minutes, and very few reached or went beyond an hour.[3] Therefore, students had ample time to complete the reading fluency and reading MSAT, as well as any of the other domains assessed.

Table 9.6.a. Descriptive statistics for cluster response time (in minutes) for non-MSAT domains

Table 9.6.b Descriptive statistics for stage response time (in minutes) for reading fluency and MSAT reading domain

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). Tables 9.7a and 9.7b show that, across all domains and up to level 4, more able students generally spent more time on each cluster or each MSAT testlet. The increase in time spent was most noticeable between students below level 1 and at level 2. However, beyond level 4, students spent slightly less time on each cluster/testlet.

Table 9.7a Median cluster response time (in minutes) by proficiency level for non-adaptive domains

Table 9.7b Median stage response time (in minutes) by proficiency level for MSAT reading

While the more able students generally needed more time to complete the test, this was not the case when response times were aggregated at the country/economy level. Figure 9.5 presents the relationship between the median total time spent on the reading MSAT items and the median reading PV1. Overall, Figure 9.5 shows that while countries/economies do vary noticeably in their average proficiency, there is no clear relationship between the average proficiency and the median total response time at the country/economy level. For example, in the case of Singapore (SGP) and Korea (KOR), both have high average reading scores, but Singapore's median response time is close to the overall median response time, while Korea's is well below it.

Figure 9.5 Median MSAT response time by median Reading proficiency score across Country/Economy

Because of differences in proficiency and other factors including motivation, the time it takes students to complete the assessment is expected to vary within each country/economy. This is shown in Figure 9.6 which presents the distribution of the total time spent on the reading MSAT items for all countries/economies, sorted by the median MSAT response time. The figure also suggests that the within-country/economy variability in response times is similar across countries/economies, regardless of the difference in languages. Since reading is the major domain for PISA 2018, and all students take a one-hour MSAT, results are presented for reading only. Nevertheless, it can be noted that the pattern for the major domain of science in PISA 2015 was similar (OECD, 2017, Figure 9.7).

[3] In a few cases, the recorded time exceeded an hour for a variety of reasons. For example, it may have been possible that the assessment "timer" was not turned off immediately upon test completion.

Figure 9.6 Distribution of reading MSAT response time in each country/economy

Item-level response time

Response time and the relationship between response time and performance were also explored at the item level. Typically, an item's median response time was highly consistent across countries/economies. For example, this is illustrated in Figure 9.7 which shows the median item response time for all GC items across countries/economies. Although there are differences across countries/economies and across languages, the pattern suggests that students generally spent more time on the items that students in other countries/economies also required a longer time to solve.

Figure 9.7 Median item response times for global competence items

Figures 9.8 and 9.9 show the median item-level response time (aggregated across all countries/economies) for the trend and new reading items, respectively, disaggregated by students' proficiency levels based on PV1. It is clear that low performing students (blue and red lines) had similar and relatively short response times across all items, while high performing students (green and purple lines) had longer response times and a larger variability in the response times across the items. This pattern was consistently observed for both the trend and new reading items.

Figure 9.8 Median item response time by proficiency level for reading trend items

Figure 9.9 Median item response time by proficiency level for new reading items

Response time reflecting possible motivation and administration issues

On average, students completed the entire test in 77.94 minutes (excluding a short break between the two assessment hours), with a standard deviation of 17.68 and a median of 80.47 minutes. Some students completed the test in less than 30 minutes (found in all countries/economies, 1.9% of the overall sample), while some students took longer than 120 minutes to complete the test (0.4% of the overall sample). At the country/economy-level, students in Albania, Colombia, Indonesia, and Malaysia took the longest time to complete the entire test, with a median time of 94.2, 89.0, 87.2, and 90.4 minutes, respectively. Students in Korea took the shortest time to complete the test, with a median time of 63.5 minutes.

There were two countries/economies where 3% or more of the students exceeded the time limit: Albania (11.5%) and Indonesia (3.3%). 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. The students with these outlier 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 the many alternative test forms made up of different clusters in different positions. For example, a student may take the reading

assessment in the first hour and then take two mathematics clusters 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 2018 main survey design balanced the order of the domains and the clusters for the PBA as well as the CBA using the BIB design. With the MSAT design, reading was delivered to students in three adaptive stages over 1 hour instead of two 30-minute clusters. Thus, position effects were also evaluated by assessment hour (i.e., 1st hour vs. 2nd hour) for all domains.

To evaluate and verify that the impact of item positions was minimal in the PISA 2018 main survey, item position effects were examined in terms of: 1) proportion of correct responses, 2) median response time, and 3) rate of omitted responses.

Table 9.8a shows the average P+ by cluster position in the CBA for the non-adaptive domains, while Table 9.8b shows the average P+ by assessment hour for all domains in the CBA. Consistent with PISA 2015, the cluster position effects were computed as the difference in P+ between position 1 and 4 (Table 9.8a). The decreases in P+ between position 1 and 4 ranged from 0.02 in mathematics to 0.07 in FL. The observed P+ position effect for the innovative GC domain was comparable to the position effects observed in the other domains. Overall, cluster position effects were very similar to the values observed in 2015. By assessment hour (Table 9.8b), 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 reading trend and new items, the decrease in average P+ between the 1st and 2nd hour was relatively small and very similar to the decreases observed in the other domains.

Table 9.8a Average proportion correct (P+) by cluster position in the CBA for non-adaptive domains

Table 9.8b Average proportion correct (P+) by assessment hour in the CBA for all domains

Tables 9.9a and 9.9b present the position effects in terms of the median response time averaged by cluster position and by assessment hour, respectively. For all domains, students spent more time on the domains when it was presented in position 1 than in position 4. FL 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 considerably more time on clusters 1 and 3, leaving them with less time for clusters 2 and 4, respectively. Table 9.9b shows the position effects by hour. As with P+, the position effects by hour were generally smaller than the position effects by cluster. For reading, the decrease in the median time between the 1st and 2nd hour was relatively small and lower than the decrease observed in mathematics, science, and FL.

Table 9.9a Median response time (in minutes) by cluster position in the CBA for non-adaptive domains

Table 9.9b Median response time (in minutes) by assessment hour in the CBA for all domains

The omission rates at different positions for all CBA countries/economies were analysed to further examine the quality of data affected by position. The omission rates are shown by cluster position and assessment hour in Tables 9.10a and 9.10b, respectively. These rates do not include the 'not-reached' items. Note that the omission rates 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 2015, the omission rate for all domains in all positions was less than 0.10, and the omission rates in positions 2 and 4 were higher than the rates in positions 1 and 3, respectively.

Table 9.10a Omission rate by cluster position in the CBA for non-adaptive domains

Table 9.10b Omission rate by assessment hour in the CBA for all domains

Position effects were also reviewed for the PBA. Tables 9.11a and 9.11b report the average P+ and the average omission rates by cluster position. With only a few countries/economies participating in the paper-based version of the assessment and a different set of countries/economies taking the PBA compared to previous cycles, some discrepancies were expected between 2018 and the previous cycles and also between the PBA and CBA results. However, no major changes from past cycles or unusual differences between the assessment modes were observed.

Table 9.11a Average proportion correct (P+) by cluster position in the PBA

Table 9.11b Average proportion of omitted responses by cluster position in the PBA

Position effects and motivation issues in the reading MSAT

Further investigations for the reading MSAT were conducted to detect potential position effects and motivational issues. For the reading MSAT, contrary to previous cycles, trend units could not be offered as part of an intact cluster (i.e., with a fixed unit order within the cluster), because the assembly of units in the testlets required a change in the unit order compared to past administrations. For this reason, the PISA 2018 field trial was specifically designed to investigate the unit order effects as a way to prepare the MSAT for the main survey (Yamamoto, Shin, Robin, Khorramdel, & Halderman, in press). More specifically, the unit order was manipulated as either fixed or variable for randomly selected group of students (see Chapter 2 for details). The field trial results confirmed the feasibility of introducing the MSAT in the main survey, as the unit order effects were found to be negligible.

To further control potential position effects, the main survey MSAT design for reading included Design B in addition to Design A by switching the order of the Stage 1 and Stage 2 testlets (see Chapter 2 for details). Therefore, position effects in the reading MSAT were further examined by stage position (Designs A and B) in addition to the assessment hour described above.

Table 9.12 shows the average percent correct, not-reached, and omission rates across all countries/economies by stage position (Designs A and B). When comparing the same testlets across Designs A and B, the results are very similar in terms of percent correct, not-reached, and omitted—for example, the percent not-reached was 1.84 for Stage 1 in design A, and the corresponding value was 1.73 for Stage 2 in design B. Comparable cells between designs A and B are indicated in grey.

Table 9.13 presents the cumulative median response time and the completion rates at each stage by stage position. Similar to the above table, the results by stage position were very similar. For both Designs A and B, the 0.2% of students who completed the MSAT at the core stage spent about 15 minutes responding to the Core items, the 4% of students who completed the MSAT at the end of stage 1 spent 43 minutes responding to the Core and Stage1 items, and the 96%

who completed the MSAT at the end of stage 2 spent about 43 minutes responding to the items in all three stages. Altogether, this confirms that Designs A and B provided comparable data.

Table 9.12 Average proportion correct, not-reached, and omitted for MSAT reading by stage position

Table 9.13 Cumulative median response time and completion rates by stage position

IRT MODELLING AND SCALING

The modelling and scaling of the PISA 2018 main survey data followed the general approach developed for PISA 2015 (OECD, 2017, Chapter 9). The following sections describe the IRT models and their assumptions, as well as the IRT scaling approach used in PISA 2018. The scaling issues associated with the new reading MSAT design and how they were resolved are addressed as well. The model-based computations developed to produce the equated proportion correct (equated P+) described earlier are also provided.

IRT models and assumptions

As in PISA 2015, the unidimensional multiple-group IRT model (Bock & Zimowski, 1997; von Davier & Yamamoto, 2004) based on the two-parameter logistic model (2PLM; Birnbaum, 1968) for the binary item responses and the generalized partial credit model (GPCM; Muraki, 1992) for the polytomous item responses were used. The 2PLM is a generalization of the Rasch model (Rasch, 1960), 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 9.1

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

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), 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 9.2

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

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

Critical assumptions of most IRT models and the models used in the 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 dependence on any demographic characteristics of the students, responses to any other items presented in a test, or the survey administration conditions. Under the unidimensional 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 this assumption, we have the formulation of the following joint probability of a particular response pattern $\boldsymbol{x}_v = (x_{v1}, \ldots, x_{vn})$ across a set of n items:

Formula 9.3

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

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 $v=1, 2, \ldots, N$) provide their answers independently of one another and that the student's proficiencies are sampled from a distribution $f(\theta)$. Using sampling weights w_v, the likelihood function is, therefore, characterised as:

Formula 9.4

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

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

[4] Note that the parametrisations $(\theta_v - b_i + d_{ir})$ 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 2PL and GPCM, representing the overal item difficulty.

an absolute scale, since any linear transformation of the item and latent trait parameters in the above formula lead to the exact same accounting of the data, often referred to as scale indeterminacy. Therefore, as part of the calibration process, a choice must be made for the IRT scale to be determined.

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

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 (OECD, 2017, Chapter 9). For each domain, a series of multi-group concurrent calibrations of the historical data (2015 and prior PISA cycles) were conducted. As a result, all the items used in all the PISA cycles up to 2015 were estimated and scaled onto the same PISA scale.

During the first run 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 (in PISA 2015, country-by-language-by-cycle group using the historic data) was evaluated. Then, item-by-group interactions were identified when the fit to the data was found to be poor (the value of the item fit statistic, discussed below, was higher than a chosen threshold value). From the second calibration run, new *unique or group-specific* item parameters were estimated in the group or groups in which misfit was found. In the subsequent calibrations, the item fit threshold was gradually lowered until the 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); Reise, Widaman, and Pugh (1993); Glas and Verhelst (1995); Yamamoto (1997); Glas and Jehangir (2014); Meredith and Teresi (2006); as well as Oliveri and von Davier (2011, 2014).

In PISA 2015 when the historic data was used, one set of common (international) parameters was estimated to model most country-by-language-by-cycle group data. Some items were allowed to have additional group-specific or unique item parameters used to model specific country-by-language-by-cycle groups (OECD, 2017, Chapter 12; von Davier et al., 2019).

The calibration and scaling for the PISA 2018 main survey followed the approach developed in 2015 and used the same IRT models. However, the historical data did not need to be included in the 2018 scaling since all trend items (reused from 2015 and/or prior PISA cycles) had already been calibrated and scaled in 2015. Therefore, in PISA 2018, a fixed item parameter linking approach was utilized with the trend item parameters fixed to their values established

in the 2015 scaling, and item parameters were estimated only for new items. Then, along with the new items, the item fit analyses of the trend items were conducted to verify whether the fixed trend item parameters are still applicable to the 2018 data and whether there is any need to re-estimate item parameters.

Item-level model-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 item 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 9.1 or 9.2 with 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 9.5

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

Formula 9.6

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

where $g = 1, \ldots, G$ is a country-by-language group; $p_g^{obs}(\theta)$ and $p_g^{exp}(\theta)$ are the observed and expected probability of correct response given proficiency θ; and $f_g(\theta)$ is the group-specific density on the students' ability scale (Khorramdel, Shin, & von Davier, 2019; von Davier, 2005). The observed probability correct is based on the pseudo counts from the EM algorithm that is used to estimate the model (Bock & Aitkin, 1981), 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 & von Davier, 2008).

The observed 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 9.1 and 9.2, RMSD is sensitive to the differences in both item difficulty and item discrimination represented by the *a* (or slope) parameter in formulae 9.1 and 9.2.

To demonstrate the use of item fit statistics (RMSD, MD), Figure 9.10 shows one example plot for a dichotomously scored item estimated via 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 that 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 9.10 Item response curve (ICC) for an item where the common item parameter is not appropriate for one group

Reading MSAT item calibration and scaling

Calibration procedures typically used with non-adaptive data may not be appropriate for MSAT data (Jewsbury & van Rijn, in press). In the reading MSAT design, Stage 1 and Stage 2 items were assigned to students based in part on their estimated proficiency, given their performance in the previous stages of the MSAT (core or core plus Stage 1) and in part based on a random process. As a result, the samples of students responding to each item are not randomly equivalent to each other but are, rather, more of less able depending on the types of testlets—easy or difficult—to which the item belongs.

In most testing situations in which MSAT is administered, item parameters are pre-calibrated (using a non-adaptive administration) and are treated as known and fixed, or at best, item fit is evaluated to examine the parameter drift (Glas, 2010). However, PISA estimates item parameters and evaluates item fit using the proper sampling weights from the national representative sample in the main survey. Calibration approaches were studied using simulation and main survey data to develop and confirm that the same scaling approach could be used for non-adaptive data implemented for the PISA 2018 minor domains would produce the desired results for the reading MSAT design.

Two important features for the reading MSAT design need to be elaborated. First, unit order effects were examined in the field trial to confirm the invariance of item parameters by unit order (Yamamoto et al., in press). If the unit order had shown to significantly impact item parameters 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. Second, in designing and finalizing the reading 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 design, 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. Furthermore, several options were considered in regards to the routing decisions to optimize adaptation (i.e., through the choice of thresholds) and to ensure the minimum exposure rate for all items (i.e., by adding the probability layer).

Such design considerations were tested and verified with simulation studies before the main survey implementation. In particular, it was shown that having the probability layer to ensure a minimum exposure rate was needed for the scaling of the reading MSAT design. Shin, Yamamoto, and Khorramdel (2018) revealed that relying solely on the student's performance on the previous set of items resulted in larger errors in the item parameter estimation[5]. For the

[5] While Jewsbury & van Rijn (in press) showed the probability layer is not necessary to ensure the item parameters were consistent, Shin et al. (2018) showed smaller errors in the item calibration with the probability layer under the PISA 2018 reading MSAT design.

main survey, the choice of the routing thresholds was based on the testlet-specific test characteristic curve (TCC) combined with the number correct (NC) score. That is, testlet-specific TCCs were calculated first and then translated into the total NC thresholds by multiplying TCC with the number of automatically scored items in each testlet. It was mainly for the straightforward computation and delivery of the test. Then, students were categorized into three performance groups in the routing procedure according to the MSAT design: low, when the total NC was less than the lower threshold for a particular testlet; medium, when it ranged between the lower and upper thresholds; and high, when it was higher than the upper threshold (see Chapter 2 Appendix for details). Different PISA scale scores were tested and evaluated for the purpose of optimizing the expected gains from the adaptive testing procedure, and the PISA scale scores of 425 and 530 were used to set the final lower threshold (between low and medium groups) and upper threshold (between medium and high groups), respectively. Results showed that the calibration of the data, simulated according to the reading MSAT design to be implemented in the main study, produced item parameters with the desired level of accuracy (Yamamoto, Shin, & Khorramdel, 2018, 2019).

The scaling procedure for the reading MSAT was further investigated using the main survey data and the simulated data. In particular, the 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. Simulation studies (van Rijn & Shin, 2019) revealed 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, Lu, & van Rijn, 2019) also provides theoretical justification for this choice.

Reading MSAT equated P+

In an MSAT design, testlets and items are assigned based largely on student performance. As a result, the subsamples of students routed to the more difficult MSAT testlets are expected to be more proficient than the total sample of students, and vice versa for students routed to the easier MSAT testlets. For this reason, CBA and MSAT classical item statistics are not always comparable. In particular, the observed proportion (or percent) correct (P+) for items included in the most difficult or easiest testlets are expected to be different from that of a sample of students randomly assigned to the items.

To facilitate comparisons between item statistics from MSAT and non-adaptive CBA designs, a new equated proportion correct statistic, or P+ EQ, has been calculated for the PISA 2018 main survey reading MSAT items. This P+ EQ is computed based on the item modelling and the mean deviation (MD) statistic described earlier. Equated proportion correct statistics are not new; they have been used by many testing programs to verify and ensure the comparability and quality of operational items and provide the information needed to assemble new test forms. Typically, they are needed when the sampled population changes over repeated test administrations or when some form of adaptive testing is employed.

The IRT model that is used to model PISA data can easily provide model-based P+, and when the model fits the data, the observed and model-based P+ for the sample of students who answered the item are expected to be the same. IRT models can also be used to accurately estimate the proportion correct for any other sample of students with known proficiency. When the sample for which the model-based P+ is produced is a reference sample, the results are called equated item proportion correct (or P+ EQ). For the dichotomously scored and partial credit items used in PISA and modelled using the 2PLM and GPCM described above, the P+ EQ can then be computed as follows (Ali & Walker, 2014):

Formula 9.7

$$P + EQ_g = \int p_g(X|\theta) f_g(\theta) d\theta, \qquad g = 1, \ldots, G.$$

$P + EQ_g$ is computed for each country-by-language group (g = 1, … ,G); $p_g(X|\theta)$ is the probability of correct response given proficiency θ; and $f_g(\theta)$ is the group-specific density distribution on the students' ability scale. To facilitate computations, the integral in the formula can be approximated by using student v sample weights w_v and summing over each student's proficiency point estimate $\hat{\theta}_v$ (WLE; Warm, 1989):

Formula 9.8

$$\int p_g(X|\theta) f_g(\theta) d\theta \approx \frac{1}{\sum_{v=1}^{N_g} w_v} \sum_{v=1}^{N_g} w_v p_g\left(X\middle|\hat{\theta}_v\right).$$

However, because PISA assesses and scores many countries/economies together on the same common scale, some relatively small degree of item model misfit may remain. MD_g defined in equation 9.5 quantifies this misfit which represents the deviation between group-specific observed and expected ICCs (von Davier, 2005). Thus, MD_g obtained as a result of the IRT scaling of the data was added in the computation of $P + EQ_g$ to account for the misfit and provide more accurate P+ EQ values:

Formula 9.9

$$P + EQ_g = \frac{1}{\sum_{v=1}^{N_g} w_v} \sum_{v=1}^{N_g} w_v p_g\left(X\middle|\hat{\theta}_v\right) + MD_g.$$

Figures 9.11a and 9.11b show examples of scatterplots of P+ versus P+ EQ for a PISA 2018 main survey participating country/economy. Items are identified by the MSAT stage (core, Stage 1 and Stage 2) and the difficulty of the testlets to which they belong (Low, Low-High, High-Low, or High). Figure 9.11a shows results obtained without using MD in the equating and Figure 9.11b shows results obtained with MD. The figures show the effectiveness of the equating and the benefit of adjusting for misfit by including MD in the computations: core item P+ and P+ EQ were nearly identical (more so in Figure b), as they should be, since no adaptation had yet taken place; Stage 1 and Stage 2 results showed the expected pattern of adjustment. That is, focusing on Figure 9.11b, we see that as Stage 1 and Stage 2 testlets are adaptively selected for more (or less) proficient subsamples of students, the items in testlets belonging to High MSAT testlets have lower P+ EQ than P+, and items in Low testlets have higher P+ EQ than P+. These patterns, as well as the Low-High or High-Low patterns (for items that belong to both Low and High testlets) that show much smaller differences between P+ and P+ EQ, are

expected given the PISA MSAT design. This confirmed the effectiveness of the equating and the choice of including MD in P+ EQ computations. Note that when the P+ values were extreme—outside of (0.025, 0.975) —the P+ EQ computations were carried out without the MD adjustment. This avoided some instances when estimates would have been negative or greater than 1. To avoid any possibility of estimates that were less than 0 or greater than 1, P+ EQ was also limited to (0, 1).

Figure 9.11a. Scatterplot of item proportion correct (P+) and equated proportion correct (P+ EQ) for one example country/economy, without adjusting P+ EQ for the country/economy item misfit MD_g

Figure 9.11b Scatterplot of item proportion correct (P+) and equated proportion correct (P+ EQ) for one example country/economy, with P+ EQ adjusted for the country/economy item misfit MD_g

Table 9.14 summarizes the effect of the MSAT design on Reading P+ and shows the average differences between P+ and P+ EQ by testlet type across countries/economies. On average, for Stage 1 and Stage 2 High testlets and Low testlets, the percent correct P+ EQ differs from the grand mean P+ by approximately 8-12%. With Stage 2 High-Low or Low-High, the MSAT effect was less than 2% on average. As expected with the core, there was no noticeable effect.

Finally, Table 9.15 shows that P+ EQ values for MS 2018 Reading trend item are consistent with the corresponding P+ values found in 2015, with a similarly strong correlation as is found for Science trend items P+ between 2015 and 2018, whereas the P+ values are less consistent (lower correlations). This further confirms that P+ EQ produces estimates that are comparable to observed P+ from previous non-adaptive PISA cycles. The same approach will be applicable in the future as the use of MSAT is extended to other domains.

Table 9.14 PISA 2018 Reading average item percent correct P+ and P+ EQ across countries/economies by MSAT stage and testlets in which the items are included

Table 9.15 PISA 2015 and 2018 correlations between Reading item P+ EQ and P+ by stage, and between science item P+'s

In short, these findings confirmed the effectiveness of the equating procedure in providing comparable proportion-correct statistics at the item-level for each country/economy across adaptive and non-adaptive designs. As expected, since core testlets were randomly assigned, the 2018 core MSAT P+ and P+ EQ were found to be equivalent and equally comparable to the P+ from 2015. Similar results were found for the Stage 2 items included in both Low-High or High-Low testlets. For Stage 1 and Stage 2 items that belonged to either high or low MSAT testlets, results clearly showed differences between P+ and P+ EQ. This was expected because the items were assigned mostly to students performing consistently high or low. Overall, the 2018 MSAT P+ EQ and 2015 CBA P+ for the trend items were found to be comparable in the same way the P+ science trend items were comparable between 2015 and 2018. As PISA expands its use of adaptive designs to other domains, this same procedure will continue to be applicable.

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, Lennon, von Davier, Gonzalez, & Yamamoto, 2013). 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; Thomas, 2002; von Davier, Gonzalez, & Mislevy, 2009; von Davier, Sinharay, Oranje, & Beaton, 2006; Wingersky, Kaplan, & Beaton, 1987).[6] 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 or three domains in a two-hour testing period. For example, in PISA 2018 for the major domain of reading, individual students responded to 33-40 items out of 245 items in total. As a result of the PISA design, a wide range of content is assessed while relatively few responses are collected from each student in any one domain in a one-hour testing duration at maximum.

Population modelling for PISA 2018 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, reading practices, academic and non-academic activities and attitudes) through multivariate latent regression models (von Davier et al., 2006). 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 regressions 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:

[6] 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.

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 9.1 and Formula 9.2 (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).
3. *Multiple imputation:* draws ten plausible values for each student on each domain from posterior distributions of proficiency using estimated Γ and Σ (Mislevy & Sheehan, 1987; von Davier, Gonzalez, & Mislevy, 2009).

Regarding Step 2, more details are provided. First, all variables in the BQ are contrast coded. 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 ($\mathbf{y}$) in addition to the item responses ($\mathbf{x}$). 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 9.10

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

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. 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 9.10 technically also depends on the item parameters, but these are treated as fixed in the computations in steps 2 and 3, and dropped from the equation. This approach is generally used to reduce computational burden ("divide-and-conquer," Patz & Junker, 1999) and motivated by the large sample sizes.

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

Formula 9.11

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

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 standard procedure entails $D = 3$ for three-dimensional (mathematics, reading, and science) model when country/economy did not choose to administer the innovative domain or $D = 4$ for four-dimensional (mathematics, reading, science, and GC) model when country/economy chose to administer the innovative domain. Latent correlations among those domains are estimated as a part of the $d \times d$ residual variance-covariance matrix. For reading subscales, overall reading is excluded and replaced with the reading subscales. For example, reading process subscales (locating information, understanding, evaluating and reflecting) are estimated based on either $D = 5$ for five-dimensional (mathematics, evaluate, locate, understand, and science) model or $D = 6$ for six-dimensional (mathematics, evaluate, locate, understand, science, and GC) model.

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

Formula 9.12

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

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), the expectation-maximization (EM) algorithm (Dempster et al., 1977) is used for maximizing the likelihood function in Formula 9.12 with respect to Γ and Σ. A multivariate variant of the latent regression model based on Laplace approximation (Thomas, 1993) is applied in reporting PISA proficiencies on more than two dimensions (domains and subdomains).

After 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 $N(\hat{\Gamma}, \widehat{V(\Gamma)})$ where $\widehat{V(\Gamma)}$ is the estimated variance of the maximum likelihood estimate $\hat{\Gamma}$ obtained from the EM algorithm (Rubin, 1987). 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 & Stout, 1993). 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) 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 9.13

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

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; Johnson & Rust, 1992; Rust, 2014) 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) 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 (9.14); 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, Maris, Bechger, & Glas, 2016; von Davier, Gonzalez, & Mislevy, 2009). 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 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 & Rubin, 1983). 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, Beaton, Kaplan, and Sheehan (1992).

After obtaining the 10 plausible values from the posterior distribution, they can be employed to evaluate the formula (9.14) for a population quantity *T* as follows:

1. Use the first vector of plausible values for each student, calculate the group estimator *T* as if the plausible values were the true values of $\boldsymbol{\theta}$. Denote the result T_1.
2. Estimate 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 (in PISA 2018 U=10), thus obtaining T_u and $V(T_u)$ for $u = 2,\ldots,U$.
4. Estimate the best estimate of the group estimator *T*, average of T_u, obtainable from the plausible values from the *U* sets of plausible values:

Formula 9.14

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

5. An estimate of the variance of group 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 $\boldsymbol{\theta}$ (often called as measurement error):

Formula 9.15

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

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 $\boldsymbol{\theta}$ are only indirectly estimated on a finite 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 9.16 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.

Table 9.16 Example for use of plausible values for partitioning the error

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 8 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 2018 MAIN SURVEY

This section describes the implementation of IRT scaling and population modelling of the PISA 2018 main survey data. The IRT scaling of each of the domains using the international data is described first. The dimensionality analyses conducted to verify the applicability of the unidimensional 2PL and GPCM models to the MSAT reading and the innovative GC 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 2018 and the prior PISA cycles is explained.

IRT scaling

IRT scaling is the first step in the modelling of PISA data. It was conducted through a multi-group IRT concurrent calibration using the international 2018 main survey data, with the trend item parameters fixed to their PISA 2015 values (common or unique country-by-language) to ensure appropriate linking to the PISA scale. Each domain was calibrated separately using the *mdltm* software (Khorramdel, Shin, & von Davier, 2019; von Davier, 2005) setup to fix already established item parameters and to estimate new ones with the unidimensional 2PL and GPCM models.

The MSAT reading and FL assessments included both trend and new items. Mathematics and science included only trend items. As the innovative domain, GC included only new items. All domains for PBA (reading, mathematics, and science) included only trend items. Table 9.17 details the number of trend items and new items by domain and mode of assessment. Note that in Reading, 245 items in total were administered, but one item (CBA DR563Q12C) had to be excluded from the analyses in all countries-by-language groups due to issues that could not be resolved. Because the PBA assessment was the same assessment as used in PISA 2015, all items were trend items and their item parameters were fixed to the values obtained from 2015 scaling. Nevertheless, the PBA 2018 data were calibrated to estimate new parameters only in cases where they no longer fit the data for particular groups or to estimate parameters for new participating countries/economies, when the common (international) parameters did not fit the data.

Table 9.17 Number of trend (linking) items and new items by domain and mode of assessment

Reading IRT calibrations were conducted using combined CBA and PBA data from 619,508 students with valid responses. For mathematics, data from 301,524 students were used; for science, data from 300,478 students were used. Only CBA data were available for FL and GC, and data from 49,660 and 75,062 students, respectively, were used in each set of IRT

calibrations. Senate weights (which sum up to 5,000 for each country/economy) were used to ensure that each country/economy contributed equally to the estimation process. 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 reading) were considered not-reached and treated as missing in the scaling analyses.

Estimation of common international and group-specific item parameters

Different language versions of the assessment used in some countries/economies could result in some items functioning differently for particular groups. These differences were considered in defining the groups specified in the multi-group IRT models. 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 the PISA 2018 main survey multiple-group IRT calibrations for reading, mathematics, and science. In FL and GC, 30 and 39 country-by-language groups were used, respectively.

To account for cultural and language differences, the stepwise calibration process described earlier was implemented to scale the 2018 data. In the first calibration and fit analyses run, for the trend items, common and group-specific item parameter estimates obtained from the PISA 2015 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[7] 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 weighted responses (based on a rescaling of the weights to sum up to 5000 within each group) 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, Khorramdel, Yamamoto, Shin, & Robin, 2019). 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 close or below RMSD of 0.18 could not be reached or cases with 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, when common item parameters showed misfit in more than one group in a similar way (direction and magnitude of the misfit), its estimation was constrained to produce the same unique parameters for this subset of groups. For example, if two groups (e.g. two countries/economies) showed poor item fit for the same item in the same direction, both groups received the same unique item parameter estimated for these two groups (note that the term *unique item parameters* in this report is used for both cases: 1) single group that receives a unique group-specific item

[7] 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.

parameter or 2) more than one group that receive the same unique item parameter that is different from the international/common item parameter). If an item showed poor fit to a different extent in different groups, different unique group-specific item parameters were used to reduce the measurement error further.

The software used for item calibration, *mdltm* readily provides the RMSD and MD fit statistics based on the formulas 9.5 and 9.6. The software implements an algorithm that monitors RMSD and MD across the specified groups and suggests a list of items to be re-estimated for a specific 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 similar misfit. Thus, the same specific item parameters may be unique to one group or multiple groups (e.g., country-by-language groups) exhibiting the similar misfit patterns. Ultimately, PISA allowed for different sets of item parameters to improve model fit and optimize the comparability of groups and countries/economies.

In most cases, the item responses across different countries/economies and language groups were accurately described by the common international item parameters. For some items, misfit led to the estimation of unique parameters for certain groups, and in some cases, the same unique parameters applied to more than one group. Outcomes of the PISA 2018 main survey analyses, including an overview of the percentage of common and group-specific item parameters across countries/economies and across PISA 2015 and 2018 main surveys, are provided in Chapter 12.

Scaling of the reading fluency items

As discussed in Chapter 2, reading fluency items were included as a part of the reading scale, which was assessed principally through the reading MSAT. These items have been introduced to increase the measurement precision at the lower level of the reading scale that was not available in prior cycles. 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, to maintain the existing reading scale and avoid any potential issues that could weaken the comparability of 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, with all the reading items parameters fixed to their final values. This approach was successfully implemented for scaling the PISA for Development Strand A data (OECD, 2018, Chapter 9). Thus, the 2018 reading assessment (reading and reading fluency) provided more information at the lower end of the scale without making substantial changes to the trend and comparability of existing reading scale.

Dimensionality analyses of the reading and global competence instruments

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, further evaluations of these assumptions are important. In particular, local dependence among items, if strong, can be addressed, and the accuracy of measurement can be improved by combining the scoring of multiple dependent items into one. The major domain of reading included new items based on the revised framework. Thus, verification that the trend and the new items are measuring the same, or very closely related, latent traits is important to ensure the comparability of proficiency over PISA cycles. For that purpose, multidimensional IRT

analyses were conducted for reading, which treated trend and new items as two different latent traits (confirmatory factor analysis). The overall model fit obtained from the two-dimensional model was found to improve only marginally over the unidimensional model, and the correlations between the unidimensional and multidimensional latent traits were very high as shown below, confirming that the use of unidimensional modelling is sufficient and appropriate. More details are provided below.

For the new innovative GC domain—where all items were new—inter-item correlations, residual and principal component analyses of field trial and main survey data and multidimensional analyses of main survey data were conducted. A residual analysis was conducted in a same way as the PISA 2015 cycle (OECD, 2017), using the response residuals computed from the scaling software, *mdltm* (von Davier, 2005).

An item response residual quantifies the difference between the model expectation and the observed item response of a respondent to an item. Using the *mdltm* software (von Davier, 2005), response residuals are computed as a follow-up step after the calibration of the items. 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 9.16

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

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

Formula 9.17

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

Formula 9.18

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

Formula 9.19

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

Once the item response residuals have been calculated, the item residual correlations across respondents can be computed to produce an item residual correlation matrix. This residual correlation is also known as the Q_3 statistic (Yen, 1984), which does not have a proper null

distribution because θ is estimated (Chen & Thissen, 1997). Unidimensional and locally independent data are expected to show random residual correlations patterns around zero across all items and across items within each unit. Local item dependencies are found when an item pair show highly correlated response residuals and their item slope parameter estimates are high. In such cases, local item dependence may be addressed by converting these two items into a single polytomous item with partial credit scoring (Rosenbaum, 1988; Wilson & Adams, 1995).

As part of the residual analysis, principal component analysis was conducted using the correlation matrix of residuals. This was to evaluate the extent to which data are unidimensional. If the unidimensional assumption holds, little common variance among the response residuals is expected after the ability dimension is accounted for.

Reading dimensionality analyses

Dimensionality analyses of the CBA reading field trial and main survey data were conducted. Because the field trial data did not lend themselves to the residual analyses described earlier, local dependencies were evaluated based on item-by-item correlations. When the local independence assumption is met, a similar level of correlation is expected among all the item pairs within a unit, and no distinctive pattern can be discerned. When exceptionally high correlation patterns among some items for a given unit were found consistently across many countries/economies, the conditional independence assumption of the IRT model could be violated. Exceptionally large slope estimates can provide similar information. In such cases, those highly correlated reading items could be combined into a single polytomous item with partial credit to remove local dependencies after discussion with content experts and item developers.

Given that the PISA items are a mixture of dichotomous and polytomous items, Spearman's rho statistic was used to estimate a rank-based measure of association. This statistic is known to be robust and has been recommended for data that does not necessarily follow a bivariate normal distribution (de Winter, Gosling, & Potter, 2016). Based on the item-by-item correlations for all reading 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. This provided evidence that the local item independence assumption was not violated.

Using the main survey data, the unidimensional IRT model described earlier showed that the trend and new items assigned to the same proficiency scale fitted the data well. The two-dimensional IRT model for the reading 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 2015 cycle, and the new items were constrained to the newly estimated unidimensional international parameters. Although AIC (Akaike, 1974) showed better fit for the two-dimensional model, BIC (Schwartz, 1978) and the log-penalty improvement showed that the unidimensional model fits better and multidimensional model provides very little improvement over the unidimensional model (Table 9.18). In particular, it was found that the unidimensional model reached 99.53% 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.

Table 9.18 Model selection criteria for the unidimensional and the two-dimensional IRT models for trend and new reading items in the main survey

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

Global competence dimensionality analyses

GC, as the innovative domain in 2018, was an entirely new domain. Preliminary classical item analyses and item-by-item correlation analyses of the GC pilot data[8] provided information about local dependencies, leading to some items to be combined and an effective scoring rule for the main survey. After the response data were rescored, residual analyses were conducted, and two additional items were combined into one polytomous item with partial credit.

For the main survey, 69 items were selected out of the 85 (partly combined) pilot study items. Unidimensional IRT scaling 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 9.12 shows the residual correlation matrix obtained. Besides the darker blue squares on the diagonal that represent each item correlating with itself, there was no other noticeable item-pair correlation pattern.

Figure 9.12 Residual correlation matrix for global competence main survey

As part of the residual analysis for the GC, a principal components analysis was conducted using the residual correlation matrix. The principal components analysis was used to further evaluate the dimensionality of the GC items. Should the eigenvalue of the first principal component extracted from response residuals be large, an additional latent trait, other than the overall ability, could be present. When residuals for all items are included as variables, the percentage of variance adds up to 100%. The percentage of variance for the first principal component ranges from 4.55% to 7.87% with a mean of 5.73%. This number can be considered a small amount of common variance. When the percentages of variance for the first 10 principal components are summed up, the value ranges from 29.12% to 45.22%, with a mean of 37.67%, a value that is more typical for a substantial amount to be considered due to a single common source of variability of response variables. The small amount of variance of the first, relative to the sum of the variances of the first ten components also shows that another dimension is not needed to explain statistical dependencies between residuals. In other words, once the ability dimension is accounted for, there is very little common variance left among the response residuals. The principal component analysis results for 6 countries/economies, as examples, are presented in Figure 9.13. Altogether, residual correlation patterns and principal component

[8] The GC items were not part of the 2018 computer-based Field Trial but were piloted on paper as part of a separate study in seven countries/economies.

analyses results provided a confirmation that the local independence and unidimensional IRT assumptions were met for the GC items.

Figure 9.13 Percentage of variance from principal component analyses for 6 countries/economies

Population modelling in PISA 2018

The population model described earlier was applied to the PISA 2018 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 GC 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 have to 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. Such derived covariates include the ratio of not-reached items, 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. Some of those derived covariates have been used in previous PISA cycles; the newly introduced changes for PISA 2018 population modelling are described in more detail in Annex H.

The following sections provide further information about how the population model was applied to PISA 2018 data, how plausible values were generated, and how plausible values can be used in further analyses.

Main sample and financial literacy sample models

The software called DGROUP (Educational Testing Service, 2012) was used to estimate the multivariate latent regression models and generate plausible values (von Davier et al. 2006; von Davier & Sinharay, 2014). 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 H 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. 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 5% of the raw sample size. This avoided a numerical instability in the estimation that could occur due to potential overparameterization of the model.

The assessment of FL was offered as an international option in PISA 2018. In total, 21 countries/economies opted to administer this assessment. The cognitive instruments included trend items from 2012 and 2015, as well as a set of new interactive items that were developed specifically for PISA 2018. Note that FL was available only in the CBA mode. Different from PISA 2015, FL was administered to a separate sample of PISA-eligible students who took a combination of reading, mathematics, and FL items. During the IRT scaling stage, the FL sample data were only used to estimate the FL item parameters, but they were not used to estimate the reading and mathematics item parameters. During the population modelling stage, the FL sample (who took Forms 73 – 84) was combined with the sample of students from the main sample who took reading and mathematics only (who took Forms 1 – 12). This was done to establish a stable linkage between the FL and main PISA forms, and the reading and mathematics domains. Thus, the FL sample received plausible values in mathematics, reading, and FL, but not in science and reading 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. A minimum of six completed cognitive items in at least one domain was considered necessary to include a student in the estimation of the population model. In PISA 2018, there were very few students[3] (0.03 %) with responses to fewer than six cognitive items in at least one of the domains assessed. 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. Nevertheless, the population model was applied to these students for the generation of plausible values. For each of the two reading subscales (by *process* and by *source*), students had to respond to at least six items in one of the subscales 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 reading) were considered not-reached and treated as missing in the population modelling.

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. To accommodate country/economy's selection of testing domains, different multivariate latent regression models were fitted on a country/economy-by-country/economy basis for the domains administered in the country/economy. For the main sample the dimensions of reading, mathematics, and science were included, together with GC, when available. For the FL, the three dimensions of FL, reading, and mathematics were included.

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

- READ for reading
- MATH for mathematics
- SCIE for science
- GLCM for GC
- FLIT for FL

Note that when assessing the correlations between domains, PVs generated from the same population model and the same draw should be used. For example, considering the main sample, each set of PVs (PV1, PV2, …, and PV10) across the mathematics, reading, science, and GC domains were drawn as a set from a multivariate distribution. Therefore, when estimating the correlations between mathematics, reading, and science, it is appropriate to use PV1 in each domain, while using PV1 in mathematics, PV2 in reading, and PV3 in science is inappropriate (because PV1, PV2, PV3 were drawn from independent multivariate conditional posterior distributions).

Population modelling for the reading subscales

The aim of generating plausible values for the different reading subscales is to provide proficiency estimates representative of important aspects within the overall reading framework. These subscales allow for secondary analyses of relationships between proficiency and BQ variables that focus of different aspects within the reading 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 reading. These were process subscales – the main cognitive process required to solve the item (locating information, understanding, evaluating and reflecting) – and source subscales – the number of text sources required to construct the correct answer to the item (single source or multiple source). Reading subscales were computed for the CBA only. Table 9.19 gives an overview of the 244 reading items by the cognitive process and the test structure. It should be noted that the two reading subscale category types are based on a two-way classification of the same 244 items (distributed into the 3+2=5 subscales). In other words, each item contributed to one of the cognitive process subscales and one of the text structure subscales.

Table 9.19 Distribution of the items to the Reading subscales

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

- Model 1: mathematics, science, GC, and the three subscales of reading cognitive process, thus, 5 or 6 dimensions in total, depending on whether GC was assessed;

- Model 2: mathematics, science, GC, and the two subscales of reading text structure subscales, thus, 4 or 5 dimensions in total, depending on whether GC was assessed.

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

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

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

- RCLI Cognitive Process Subscale of Reading – Locating Information
- RCUN Cognitive Process Subscale of Reading – Understanding
- RCER Cognitive Process Subscale of Reading – Evaluating and Reflecting
- RTSN Source Subscale of Reading – Single source
- RTML Source Subscale of Reading – Multiple source

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 RCLI1, RCN1, RCER1 is appropriate, while estimating correlations between RCLI1, RCN2, RCER3 is inappropriate. The same is true for the source subscale. Because the core and innovative domain PVs and the subscore 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 12.

Linking PISA 2018 to previous PISA cycles

PISA accounts for measurement errors due to student sampling, the reliability of the assessment, and 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. This variability over time and over different PISA assessment designs (minor/major, etc.), independent item calibration based on the subset of samples for each cycle at a time, as well as the fact that we do not "know" the true difficulty and discrimination of items, introduces a source of uncertainty in the results. It becomes apparent as soon as there are multiple samples that were collected and calibrated on the subset of samples successively that the item parameter estimates tend to be slightly different every time new data is collected and these parameters are calculated. This, in turn, has an effect on the results reported to countries/economies, and it can be quantified in the linking error. This linking error is a part of the variability of country/economy means that is due to the tests not being exactly the same and having different samples of students in the estimation of item parameters.

In summary, the uncertainty due to linking can result from changes in the assessment design or the scaling procedure used, such as:

1. different calibration samples used to estimate parameters in different cycles,
2. usage of entire sampled data instead of subsamples of data,
3. the inclusion of items that are unique to each cycle in addition to common items,
4. changes in the cluster position within the assessment (PISA 2000 was an unbalanced design; later designs balanced cluster positions except for PISA 2018 MSAT for reading),
5. changes in the model used for scaling, and
6. the particular set of trend items that are common between assessment cycles of interest and which can be seen as one among an infinite set of possible trend items.

In PISA, it is important to note that the composition of the assessment in any two cycles are different due to changes in the domain emphasis, recombination of items and units within clusters, framework changes, assessment mode changes, and test design changes. Although the reporting model remains a unidimensional IRT model, which fits quite well, trend items are modelled based on data collected in different contexts. Thus, estimating linking error for trend measures is important to account for cycle-to-cycle differences.

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 variance estimation based on replicate weights and plausible values and replicate weights-based. 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 9.21

$$\tilde{\theta}_{v,C1,g} = \theta_{v,true} + \hat{u}_{C1,g} + \tilde{e}_v, \tag{9.23}$$

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 9.22

$$\tilde{\theta}_{v,C2,g} = \theta_{v,true} + \hat{u}_{C2,g} + \tilde{e}_v. \tag{9.24}$$

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 9.23

$$\tilde{\theta}_{v,C1,g} - \tilde{\theta}_{v,C2,g} = \hat{u}_{C1,g} - \hat{u}_{C2,g}, \tag{9.25}$$

and the expectation can be estimated by:

Formula 9.24

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

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 moments. That is, we may assume:

Formula 9.25

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

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 link error can be written as:

Formula 9.26

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

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 (9.26) 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. The S_n statistic was proposed by Rousseeuw and Croux (1993) 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 9.27

$$S_n = 1.1926 * \text{med}_i \left(\text{med}_j \left(|x_i - x_j| \right) \right).$$

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 12.

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*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Ali, U. S., & Walker, M. E. (2014). *Enhancing the equating of item difficulty metrics: Estimation of reference distribution* (Research Report No. RR-14-07). Educational Testing Service. https://doi.org/10.1002/ets2.12006

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). Springer-Verlag.

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

de Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, *21*(3), 273–290. https://doi.org/10.1037/met0000079

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

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. Springer.

Glas, C. A. W. (2010). Item parameter estimation and item fit analysis. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 269–288). Springer.

Glas, C. A. W., & Jehangir, K. (2014). Modelling country specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 97–115). CRC Press.

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). Springer.

Jewsbury, P., Lu, R., & van Rijn, P. W. (2019). *Modeling multistage and targeted testing data with item response theory* [Manuscript submitted for publication]. Research and Development Division, Educational Testing Service.

Jewsbury, P. A., & van Rijn, P. W. (in press). Item calibration in multistage tests. In D. Yan (Ed.), *Research for practical issues and solutions in computerized multistage testing*. Chapman & Hall.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics*, *14*(4), 303–334. https://doi.org/10.3102/10769986014004303

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*, *17*(2), 175–190. https://doi.org/10.3102/10769986017002175

Joo, S. H., Khorramdel, L., Yamamoto, K., Shin, H. J., & Robin, F. (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.

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM software mdltm including parallel EM algorithm. In M. von Davier & Y. S. Lee (Eds.), *Handbook of psychometric models for cognitive diagnosis* (pp. 603–628). Springer.

Kirsch, I., Lennon, M. L., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1–11). Springer. https://doi.org/10.1007/978-94-007-4629-9_1

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4). 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data. *American Statistician*, *37*(3), 218–220.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from plausible values? *Psychometrika*, *81*(2), 274–289. https://doi.org/10.1007/s11336-016-9497-x

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

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

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

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R. J. & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (pp. 361–380). Educational Testing Service.

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

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modelling*, *53*(3), 315–333.

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*(1), 1–21. https://doi.org/*10.1080/15305058.2013.825265*

Organisation for Economic Co-Operation and Development. (2017). *PISA 2015 technical report*. *https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf*

Organisation for Economic Co-Operation and Development. (2018). *PISA for Development technical report*. https://*www.oecd.org/pisa/pisa-for-development/pisafordevelopment2018technicalreport*

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178. https://doi.org/10.3102/10769986024002146

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen and Lydiche.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Rosenbaum, P. R. (1988). Permutation tests for matched pairs with adjustments for covariates. *Applied Statistics*, *37*(3), 401–411. https://doi.org/10.2307/2347314

Rousseeuw, P. J. & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283. https://doi.org/*10.2307/2291267*

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley and Sons.

Rust, K. F. (2014). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 117–154). CRC Press.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shin, H. J., Yamamoto, K., & Khorramdel, L. (2018). *Increasing the measurement efficiency and accuracy of PISA through multistage adaptive testing design* [Manuscript in preparation]. Research and Development Division, Educational Testing Service.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, *2*(3), 309–322.

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

van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1–28). Springer.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (2016). *Handbook of modern item response theory* (2nd ed.). Springer.

van Rijn, P. W., & Shin, H. J. (2019). Item calibration for multistage tests in the context of large-scale educational assessment [Manuscript in preparation]. Research and Development Division, Educational Testing Service.

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

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI monograph series, 2(1), 9–36. http://www.ierinstitute.org/IERI_Monograph_Volume_02_Chapter_01.pdf*

von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). CRC Press.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. E. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 1039–1055). Elsevier. https://doi.org/10.1016/S0169-7161(06)26032-2

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

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

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*(2), 181–198. https://doi.org/10.1007/BF02301412

Wingersky, M., Kaplan, B., & Beaton, A. E. (1987). Joint estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (pp. 285–292). Educational Testing Service.

Xu, X., & von Davier, M. (2008). *Comparing multiple - group multinomial log - linear models for multidimensional skill distributions in the general diagnostic model* (ETS Research Report Series ETS RR-08-35). Princeton, NJ: Educational Testing Service.

Yamamoto, K. (1997). Scaling and scale linking. In T. S. Murray, I. S. Kirsch, & L. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey* (pp. 161–178). National Center for Education Statistics.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, *37*(4), 16–27. https://doi.org/10.1111/emip.12226

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). *Introduction of multistage adaptive testing design in PISA 2018* (OECD Education Working Papers No. 209). OECD Publishing. https://doi.org/10.1787/b9435d4b-en

Yamamoto, K., Shin, H. J., Robin, F., Khorramdel, L. & Halderman, L. (in press). Improved test designs and multistage adaptive testing in large-scale assessments. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative computer-based international large-scale assessments: Foundations, methodologies and quality assurance procedures*. Springer.

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