Chapter 12
Scaling outcomes

This chapter reports on the outcomes of implementing the item response theory (IRT) and population modelling methods described in Chapter 9 for the PISA 2018 main survey cognitive assessment data. It provides assessments of the invariance of the IRT item parameters across countries/economies, estimates of the reliability and correlations across assessments domains, and estimates of the linking errors between the 2018 and prior PISA cycles. The charateristics of the item pool are evaluated, and the items are classified across proficiency levels based on their common international parameters. Finally, the correlations between scales and the percentage of students in each country at each proficiency level are presented for each cognitive domain.

RESULTS OF THE IRT SCALING AND POPULATION MODELLING

Results of the IRT scaling and population modelling include descriptions of the proportions of item parameters that were invariant across countries and PISA cycles, as well as the reliability of the cognitive assessments for each country/economy. The following sections illustrate that the comparability of the PISA scales across cycles and countries was achieved in each domain by reaching a desirable proportion of invariant item parameters across countries/economies and cycles.

Assessing the invariance of item parameters

The item parameters for all the items used in the assessment were obtained through IRT scaling. In PISA 2018, IRT scaling was implemented through a multi-group (i.e., country-by-language groups) IRT concurrent calibration using the 2018 main survey data, using the trend items as fixed linking items and setting the scale to the PISA scale established in 2015. That is, item parameters for trend items were fixed to the ones used in PISA 2015 (either common international or unique to a specific country-by-language group or groups), unless there was evidence that the 2015 parameters did not fit the 2018 data (see Chapter 9 for details).

In most cases the international item parameters fitted data for all country-by-language groups. When they did not fit a particular country-by-language group, unique or group-specific parameters were computed, and if no parameters could be found that fit data in the country-by-language group or groups, the item was dropped for these groups. In total, only one reading item (DR563Q12C) was identified as problematic based on classical item analyses, and the IRT parameters did not fit the data for the majority of the country-by-language groups. Feedback from countries and further content review showed the item to be flawed, so it was dropped from all groups.

To assess the invariance of item parameters across country-by-language groups and cycles, items were categorized as:

- *invariant* when common international parameters could be used;
- *group-specific invariant* when the same unique parameters could be used across cycles (applies only to trend items);

- *variant* for all other cases where unique item parameters were estimated (new items) or when unique parameters were estimated that are different from the 2015 parameters (trend items); and
- *dropped* when the item could not be fitted to the data and was dropped for one or more country-by-language groups.

For countries with multiple language groups, the number of invariant, variant, or dropped items were averaged across the different language groups within the country to calculate the proportion of unique item parameters used. Sample weights were used for this calculatiton.

Table 12.1 shows the proportions of items categorized as invariant, variant, and dropped, averaged across countries participating in the 2018 computer-based assessment (CBA). The proportion of invariant items was large for all domains, ranging from 77.1% for the reading trend items to 95.2% for the reading fluency items. A large proportion of invariant items is critical for ensuring the comparability of scores across countries and cycles. Group-specific invariant items also contribute to the comparability of scores across cycles. The proportion of invariant total (invariant and group-specific invariant) was near or above 90% for all domains. Regarding the dropped category, the proportions were very small for all domains (0.6% or less).

Table 12.1 Proportion of invariant, variant, and dropped CBA items averaged across countries/economies, for each domain

Table 12.2 shows the proportion of items categorized as invariant, variant, and dropped, averaged across countries participating in the 2018 paper-based assessment (PBA). The results are similar to those for CBA, with: high proportions of invariant items ranging from 75.2% in reading to 88.2% in mathematics; high proportions of invariant total items ranging from slightly below 90% for reading and above 90% for mathematics and science and slightly below 90% for reading; and very small proportions of dropped items.

Table 12.2 Proportion of invariant, variant, and dropped PBA items averaged across countries/economies, for each domain

An overview of the frequencies of invariant, variant, and dropped items for each domain is presented in Figures 12.1 to 12.5 for CBA and PBA participating countries. Each country is represented by a vertical bar: Dark green represents the number of items classified as invariant (for reading and financial literacy, a vertical bar is used to separate the trend and new items, with trend items clustered around the x-axis and new items at each end of the bar); light green represents the number of group-specific invariant items (only trend items); yellow represents the number of variant items[1] (a horizontal bar separates the trend and new items in reading and financial literacy); and red represents the number of items dropped from scaling. The variant and dropped items frequencies are shown using negative values to highlight differences between the number of items that contribute to ensuring the comparability of the PISA scales

[1] For the trend items classified as variant in a specific group (yellow), the 2015 parameters did not appropriately fit the 2018 data; thus, new unique parameters were estimated. For new items classified as variant in a specific group (yellow), unique parameters were needed due to the misfit of the common international parameters to the 2018 data.

(invariant) and the number of items that do not. The countries are sorted from left to right by increasing number of invariant items.

These plots show that while there is some variability across countries, the numbers of invariant item parameters and group-specific invariant item parameters are large.

Figure 12.1 Frequency of invariant, variant, and dropped items for reading, by country/economy

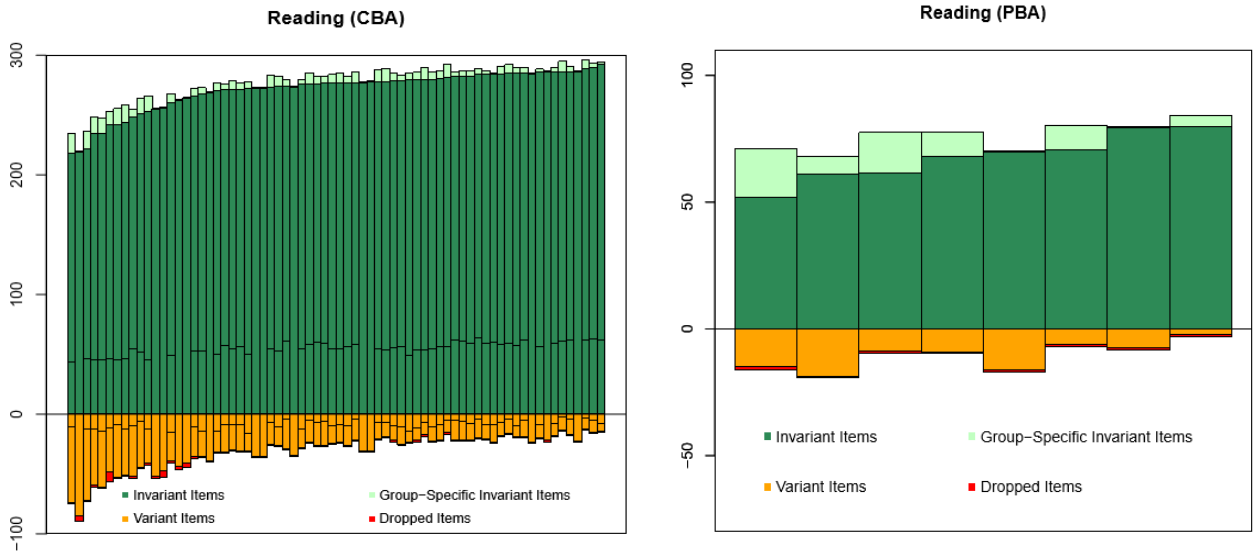


Figure 12.2 Frequency of invariant, variant, and dropped items for mathematics, by country/economy

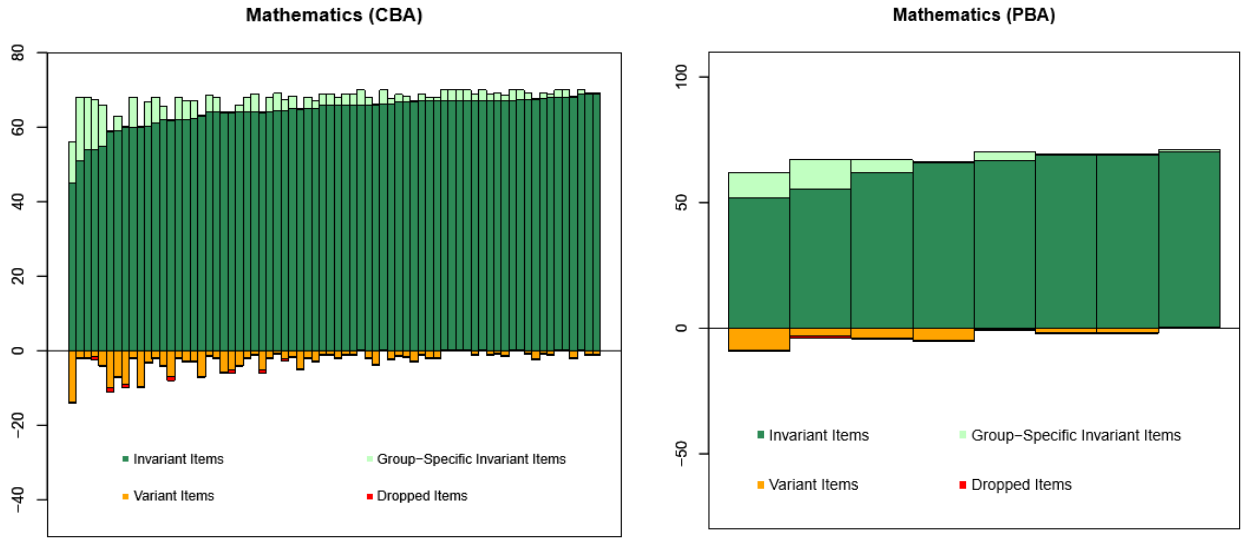


Figure 12.3 Frequency of invariant, variant, and dropped items for science, by country/economy

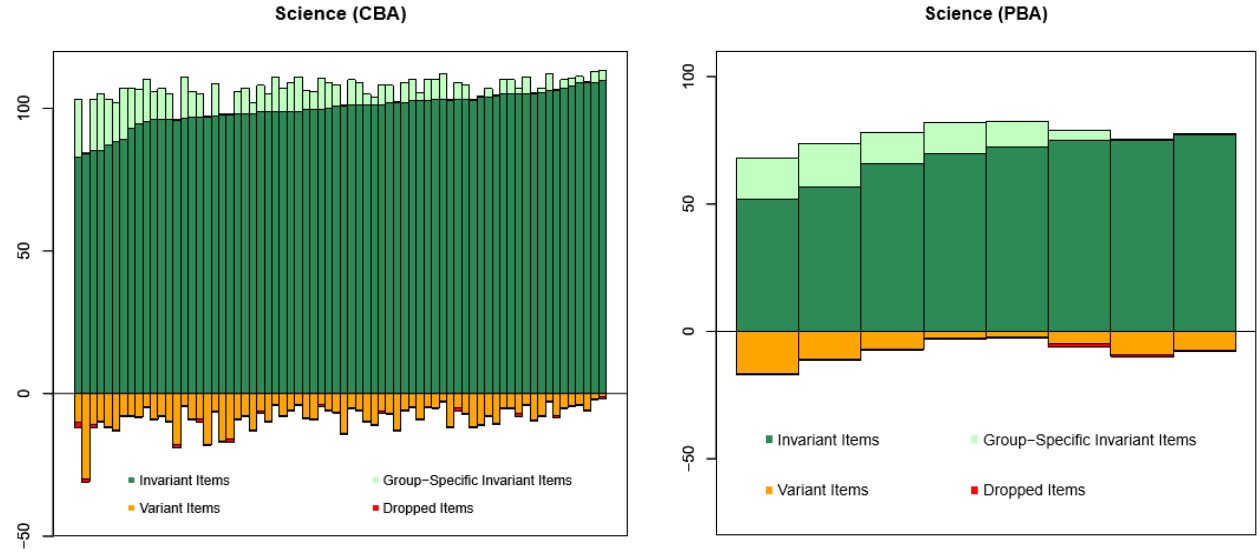


Figure 12.4 Frequency of invariant, variant, and dropped items for financial literacy, by country/economy

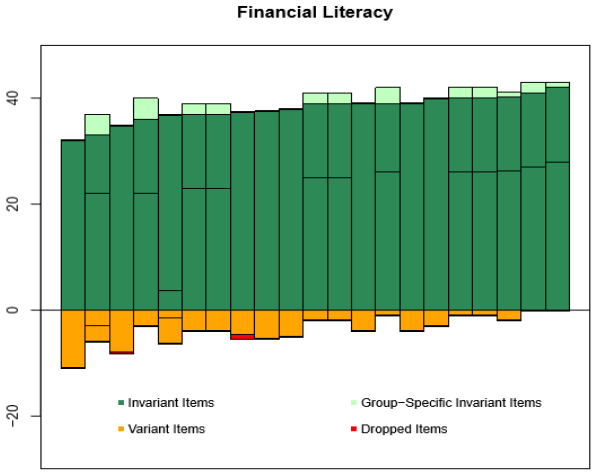


Figure 12.5 Frequency of invariant, variant, and dropped items for global competence, by country/economy (Forthcoming on 22 October 2020)

After the IRT scaling was finalised, item parameter estimates were delivered to each country, with an indication of which items had received common international item parameters and which had received unique group-specific item parameters. Table 12.3 gives an example of the information provided to countries: The first column shows the domain, the second column flags items that had received group-specific parameters (unique item parameters) or had been excluded from the IRT scaling (excluded from scaling) with empty cells indicating invariant item parameters were retained, and the remaining columns show the final item parameter estimates (the slope and difficulty parameters are listed for all items, while the threshold parameters are listed for the polytomous items). All new items in the 2018 main survey data were scaled using the two-parameter logistic model (2PLM; Lord and Novick, 1968) or the generalized partial credit model (GPCM; Muraki, 1992).

Table 12.3 Example of item parameters estimates provided to countries/economies showing item-by-country level treatment

Reliability of the PISA scales

Plausible values were generated for all students by fixing all item parameters to the values obtained from the final IRT scaling and by fitting the multivariate latent regression models described in Chapter 9.

Given the multi-stage adaptive testing (MSAT) for reading domain and rotated and incomplete assessment design for other domains, it was not possible to calculate the classical reliability values for each cognitive domain. Nevertheless, test reliability was estimated in the same way it was in PISA 2015, using the commonly used formula: *1 – (expected error variance/total variance)*. The expected error variance is the weighted average of the posteriori variance (i.e. the variance across the 10 plausible values, which is an expression of the posterior measurement error). The total variance was estimated using a resampling approach (Efron, 1982) and was estimated for each country depending on the country-specific proficiency distributions for each cognitive domain.

Table 12.4 presents the distribution of the national reliabilities for the generated scale scores calculated with all 10 plausible values. The reliabilities for each country are presented in Table 12.5. These tables show that the variance explained by the combined IRT model and population model is comparable across countries. While the median values are above 0.80 in all the domains assessed in CBA and PBA, it is important to keep in mind that this is not to be confused with a classical reliability coefficient, as it is based on more than the item responses.

Table 12.4 Descriptive statistics of the national reliabilities of the cognitive domains and reading subscales

Table 12.5 National reliability values of the cognitive domains

Reading MSAT measurement error

As indicated earlier, the main goal of the new reading MSAT design was to improve measurement precision over what would have been obtained with the linear (non-adaptive) design used in past PISA cycles. A simulation study based on the item parameters obtained from the field trial and MSAT design for reading showed that measurement precision was expected to increase by as much as 10% at the lower and higher proficiency levels (Yamamoto, Shin & Khorramdel, 2019). Results of the reading MSAT implemented in the 2018 main survey confimed these expectations.

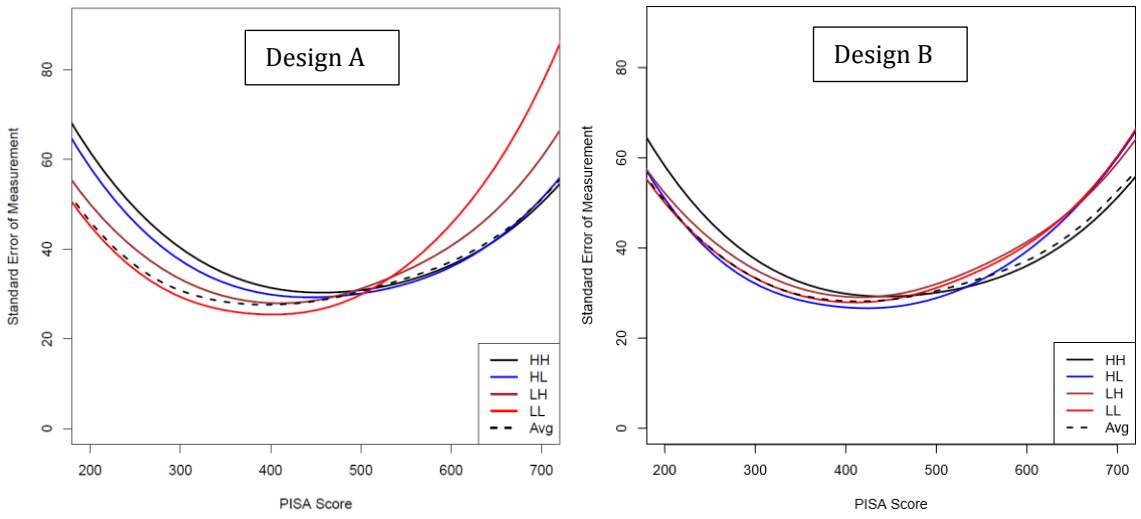
Using the common international item parameters estimated from the IRT scaling of the PISA 2018 main survey data, the MSAT average standard errors of measurement were computed across the full range of PISA reading scale from 200 to 700. More specifically, for each path, the standard error of measurement was computed as the reciprocal of square root of the information function, scaled on the PISA reporting scale: $SE(\theta) = A * 1/\sqrt{I(\theta)}$, where A is the transformation coefficient provided in Table 12.6 below.

Then, the path-specific standard errors were averaged across the testlet combinations: core and difficult testlets in both stages (HH), core and easy testlets in both stages (LL), core, a difficult

testlet in stage 1, and an easy testlet in stage 2 (HL), and core, an easy testlet in stage 1, and a difficult testlet in stage 2 (LH).

The results are displayed in Figure 12.6 for designs A and B (see Chapter 2 of this Technical Report for details about these designs). The lowest standard error of measurement (i.e., highest measurement accuracy) that could be achieved is shown by the lowest point across the the PISA scale—if all students were routed according to their true proficiency. For example, looking at design A, for a student with a true proficiency of 350, the easiest (LL) path provides the lowest standard error of measurement (approximately 30 points). For a student with a true proficiency of 700, the most difficult (HH) path provides the lowest standard error of measurement (approximately 50 points). However, some students were "misrouted" to less informative paths because the proficiency estimates used in routing contain measurement error, and a small proportion of students were routed randomly to ensure a minimum exposure rate for all items and proper coverage of the contents across the entire population of students. Therefore, the average standard error of measurement obtained at each point on the PISA scale (i.e., black dashed line in Figure 12.6) was calculated as the average of the standard errors of measurement for each HH, HL, LH, and LL path, weighted by the proportion of sudents routed to each path. As a result, the weighted average standard errors of measurement are somewhat higher than the lowest possible values, but very close to the lowest possible values. That is, the weighted average (dashed line) is very close to the LL (red line) at the lower level and to HH and HL (blue and black lines) at the higher level. Despite the intentional misrouting of some students, results showed that that the adaptation worked well at all proficiency levels, as the weighted average is only slightly above the lowest standard error of measurement that can be achieved across the different MSAT paths. A comparison of the Designs A and B also shows that Design B is less adaptive than Design A because its paths are less differentiated in difficulty and information and its routing less accurate.

Figure 12.6 Reading MSAT conditional standard error of measurement by form (HH, HL, LH, LL) and in weighted average across the actual assigned forms for MSAT Designs A and B

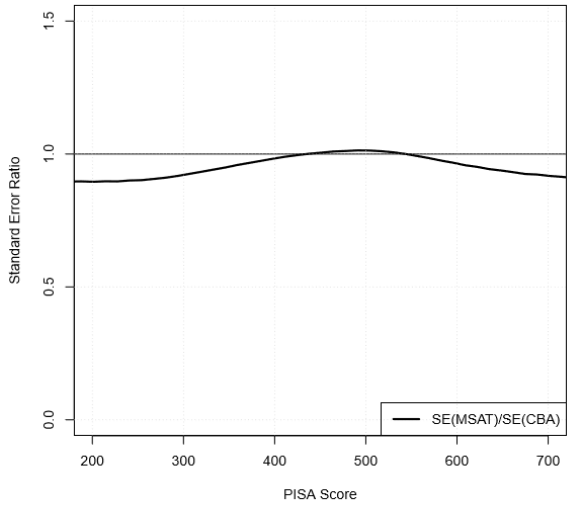


To assess the efficiency of the reading MSAT designs, the average standard error of measurement of the reading MSAT design, A or B, assigned to students was compared to a corresponding non-adaptive PISA design (CBA) that could have been implemented using the same items and the same average test length. The CBA standard errors of measurement across the PISA scale were computed in a similar way as with the MSAT—as the average of the item

standard errors over all MSAT items multiplied by the expected number of items across all the MSAT possible paths (36.7 items).

Figure 12.7 shows the ratio of the standard errors of measurement for the reading MSAT design that was implemented in the 2018 main survey to the corresponding non-adaptive design that could have been implemented. A ratio of MSAT to CBA of less than 1 indicates that the standard error of measurement for the MSAT design is lower than that of the corresponding non-adaptive CBA design. Thus, the implemented MSAT reduced the standard error of measurement by as much as 10% at the lower and higher proficiency levels—which is close to the results found based on simulated data (Yamamoto, et al., 2019).

Figure 12.7 Ratio of the standard errors of measurement for the reading MSAT design (A and B combined) to the standard errors of measurement for a corresponding non-adaptive CBA design



TRANSFORMING THE PLAUSIBLE VALUES TO PISA SCALES

The plausible values generated from the population latent regression models need to be transformed using a linear transformation. This set of plausible values transformed to the PISA scales can then be used to compare the overall performance of countries or subgroups within a country or across cycles.

Reading, mathematics, and science

For reading, mathematics, and science, the transformation coefficients established in the PISA 2015 cycle were applicable to the 2018 cycle because the 2018 IRT parameters were estimated to be on the same scale as the 2015 item parameters. Note that in 2015, the transformation coefficients were computed for each domain based on the 2006, 2009, 2012, and 2015 scaled proficiencies from only the OECD countries. The country means and variances used to compute the transformation coefficients included only the values from the cycle in which a given content domain was the major domain. Hence, the transformation coefficients for science are based on the 2006 reported results, the reading coefficients are based on the 2009 results, and the mathematics coefficients are based on the 2012 results. Computational details are provided in the PISA 2015 technical report (OECD, 2017, Chapter 12).

Financial literacy

For financial literacy, results from the 2012 PISA cycle were used to compute the transformation coefficients. The approach used for computing the transformation coefficients was the same as used for reading, mathematics, and science. However, all available country data were used to compute the financial literacy coefficients, whereas for reading, mathematics, and science, only the data from OECD countries were used. This decision was made because there were relatively few OECD countries that had participated in the financial literacy assessment, and using all countries provided transformation coefficients and a scale that was more appropriopriate for all participating countries.

Global competence

Global competence was a newly established domain in PISA 2018. Consistent with processing of new domains that had been introduced in previous PISA cycles, the transformation coefficients for global competence were computed so that the plausible values for the 10 OECD participating countries would have a mean of 500 and a standard deviation of 100. To take into account the 10 sets of plausible values, all sets were stacked together and the weighted mean and variance were computed so that each country contributed equally. As a result, the full set of transformed plausible values for global competence had a weighted mean of 500 and a weighted standard deviation of 100 for the OECD countries.

Specifically, the equations used to compute the transformation coefficients for global competence are presented below. In formula 12.1, w_v is the senate weight for for student v $\{v=1, 2, \ldots, N\}$ and X_{vu} is the u^{th} plausible value $\{u = 1, 2, ..., 10\}$ for student v. The weighted grand mean across all 10 plausible values is $\bar{X}$, which is computed by compiling all 10 sets of plausible values into a single vector (with the corresponding senate weights compiled in a separate vector) and finding the weighted mean of these values. The weighted variance of the plausible values is τ_{PV}^2, which is computed using the vector of plausible values described above. The square root of τ_{PV}^2 is the weighted standard deviation, τ_{PV}.

Formula 12.1

$$\tau_{PV} = \sqrt{\tau_{PV}^2} = \sqrt{\frac{\sum_{u=1}^{10}\sum_{v=1}^{N} W_v(X_{vu}-\bar{X})^2}{[(10N-1)\sum_{v=1}^{N} w_v]/N}}$$

The transformation coefficients for global competence were computed using the following equations:

Formula 12.2

$$A = \frac{100}{\tau_{PV}}$$

Formula 12.3

$$B = 500 - A[\bar{X}] = 500 - A\left[\frac{\sum_{u=1}^{10}\sum_{v=1}^{N} w_v X_{vu}}{10\sum_{v=1}^{N} w_v}\right]$$

The plausible values for global competence were transformed to the PISA scale using a similar approach to that used for reading, mathematics, science, and financial literacy. However, one difference is that, for global competence, the transformation was based solely on the 2018 plausible values because global competence was introduced for the first time in 2018.

Transformation coefficients for all domains

The transformation coefficients for all content domains are presented in Table 12.6. The A coefficient adjusts the variability (standard deviation) of the resulting scale, while the B coefficient adjusts the scale location (mean).

Table 12.6 Transformation coefficients for PISA 2018

Table 12.7 shows the average transformed plausible values as well as the resampling-based standard errors for each country and domain.

Table 12.7 Average plausible values (PV) and resampling-based standard errors (SE) by country and domain

LINKING ERROR

The estimation of the linking error between two PISA cycles was accomplished by considering the differences between the reported country means from the previous PISA cycles and new estimates of these country means based on the new PISA cycle item parameters. To estimate the linking error for trend comparisons between PISA 2018 and a previous PISA cycle, the subset of countries that had participated in both cycles being compared was used. In the case of financial literacy, since the number of participating countries was relatively small, all countries were used.

The 2018 linking errors are reported in Table 12.8 below. Using these values help evaluate the extent to which changes in a country/economy or subgroup's performance between PISA 2018 and a previous PISA cycle are significantly different.

Note that for each domain, the earliest cycle for which comparisons can be made between PISA 2018 and a previous PISA cycle is the cycle in which the domain first became a major domain. Thus, the comparison of mathematics scores between PISA 2018 and PISA 2000 is not possible, nor is the comparison of science scores between PISA 2018 and PISA 2000 or between PISA 2018 and PISA 2003.

Table 12.8 Linking error for score comparisons between PISA 2018 and previous PISA cycles

INTERNATIONAL CHARACTERISTICS OF THE ITEM POOL

This section provides an overview of the test targeting, the domain inter-correlations, and the correlations among the reading scale and subscales.

Test targeting

Similar to assigning a specific score on a scale to students according to their performance on an assessment (OECD, 2002), each item in PISA 2018 was assigned a specific value on a scale

based on response probability (RP) calculated using the item's IRT parameters (discrimination and difficulty). Chapter 15 describes how items can be placed along a scale based on their RP values and how these values can be used to classify items into proficiency levels.

In PISA, the RP62 values were used to classify items into levels. Students with a proficiency located at or below this point have a 62 percent or less probability of getting the item correct, while students with a proficiency above this point have a higher than 62 percent probability of getting the item correct. The RP62 values for all items are presented in Annex A, together with the final item parameter estimates obtained from the IRT scaling.

Similar to the process above, students were also classified into proficiency levels using the plausible values. For each cognitive domain, the levels were defined by equidistant score boundaries which were determined based on the previous PISA cycles. Tables 12.9 to 12.13 show the levels defined for for each cognitive domain, along with the percentage of items and students classified at each level of proficiency.

Table 12.9 Proficiency levels for reading and the classification of items and students

Table 12.10 Proficiency levels for mathematics and the classification of items and students

Table 12.11 Proficiency levels for science and the classification of items and students

Table 12.12 Proficiency levels for financial literacy and the classification of items and students

Table 12.13 Proficiency levels for global competence and the classification of items and students (Forthcoming on 22 October 2020)

Since RP62 values and the plausible values are on the same scale, the distribution of students' latent ability and the items' RP62 values can be compared and contrasted. In Figures 12.8 to 12.12, the left side of each figure illustrates the distribution of the first plausible values (PV1) across countries. In each figure, the blue line indicates the empirical density of the first plausible values across all countries, and the red line indicates the theoretical normal distribution with the mean and the variance of plausible values across all countries in each domain. The figures show that the distribution of the plausible values for each domain are approximately normal. On the right side of each figure, the RP62 value for each of the items is plotted. For polytomous items, only the highest category's RP62 value is used, and this was the value used to assign the item to a proficiency level.

Figure 12.8 Distribution of the first plausible values and item RP62 values in reading

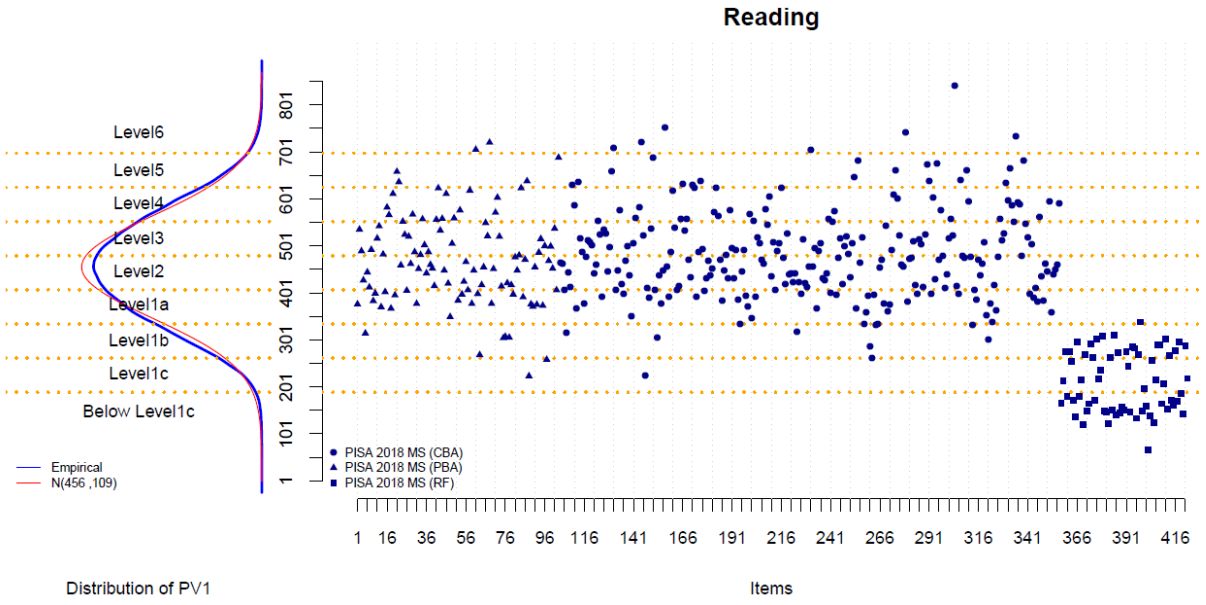


Figure 12.9 Distribution of the first plausible values and item RP62 values in mathematics

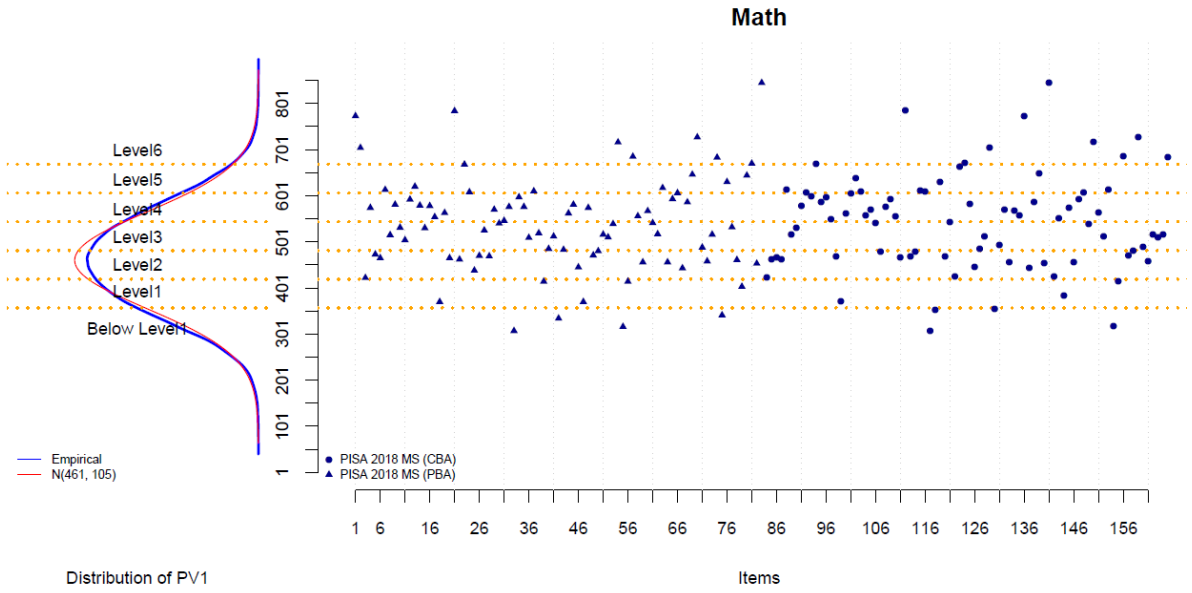


Figure 12.10 Distribution of the first plausible values and item RP62 values in science

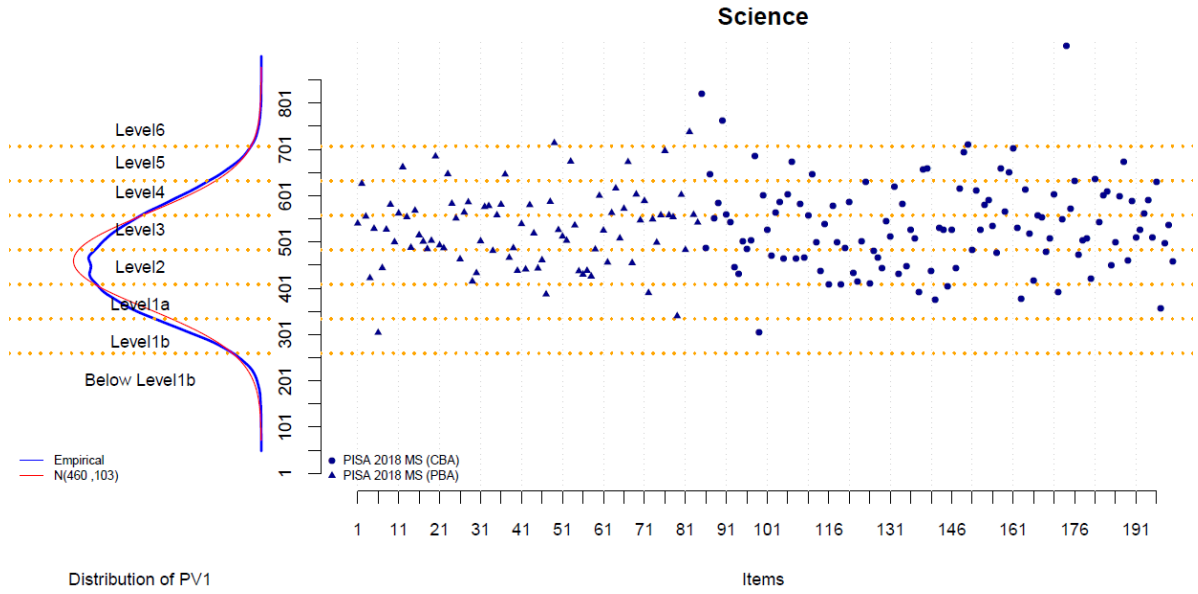


Figure 12.11 Distribution of the first plausible values and item RP62 values in financial literacy

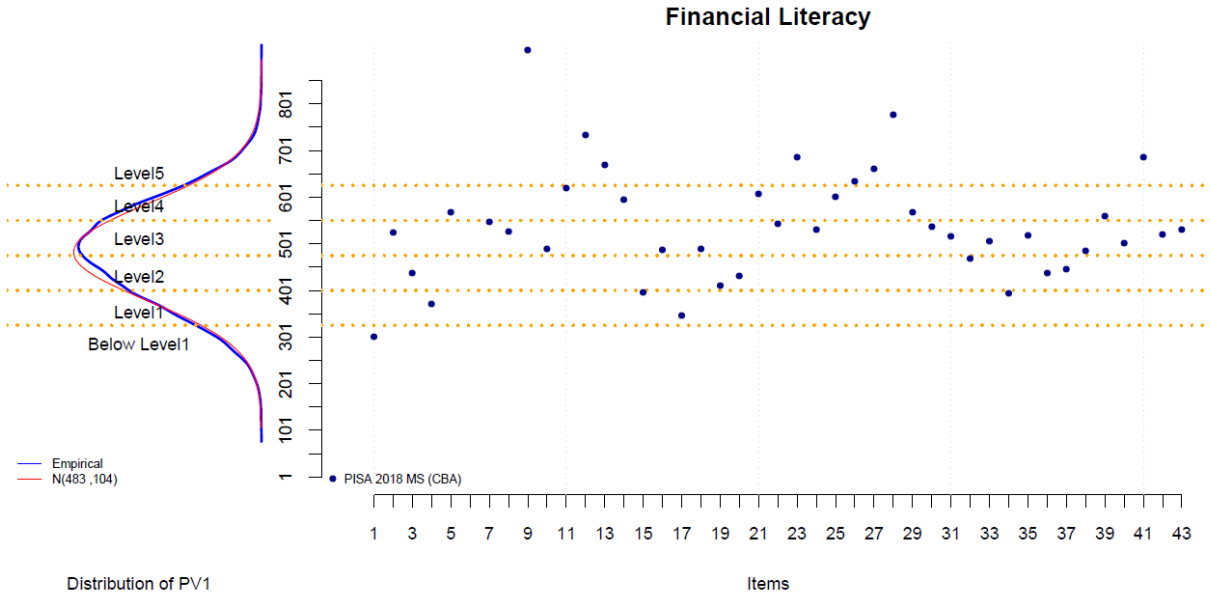


Figure 12.12 Distribution of the first plausible values and item RP62 values in global competence (Forthcoming on 22 October 2020)

DOMAIN INTER-CORRELATIONS

Estimated correlations between the domains, based on the 10 plausible values and averaged across all countries and assessment modes, are presented in Table 12.14 for the main sample

and in Table 12.5 for the financial literacy sample. The estimated correlations for each country are presented in Table 12.16.

Table 12.14 Domain inter-correlations for the main sample

Table 12.15 Domain inter-correlations for the financial literacy sample

Table 12.16 Domain inter-correlations by country/economy

Correlations with the Reading subscales

There were two sets of subscales reported for reading. The first set, measuring cognitive processes, was composed of the following subscales: evaluating and reflecting (RCER), locating information (RCLI), and understanding (RCUN). The second set, based on the text source, comprised the subscales multiple source (RTML) and single source (RTSN).

The correlations between the cognitive domains and the cognitive processes reading subscales are presented in Table 12.17. The correlations between the cognitive domains and the text structure reading subscales are presented in Table 12.18.

Note that, as indicated in Chapter 9, because of the way in which these subscale plausible values were estimated, it is not appropriate to correlate the cognitive process subscales with the text source subscales, or any of the subscales with the overall reading proficiency.

PERCENTAGE OF RESPONDENTS AT EACH PROFICIENCY LEVEL

Figures 12.13 to 12.17 show the percentage of students in each country at each proficiency level for each cognitive domain.

REFERENCES

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans.*CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38*. http://dx.doi.org/10.1137/1.9781611970319

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

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

Organisation for Economic Co-Operation and Development (2002). *Reading for Change: Performance and Engagement across Countries (Results from PISA 2000).* http://dx.doi.org/10.1787/9789264099289-en.

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

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