Chapter 10: SCALING AND POPULATION MODELLING METHODS FOR COGNITIVE DATA

INTRODUCTION

This chapter describes the quantity and quality of the data submitted by the participating countries. Analyses were conducted to verify whether the data had been collected according to the test design and whether the data quality was appropriate to apply the scaling and population modelling methods. In the following sections, the models and methods used for the item response theory (IRT) scaling, population modelling, and the generation of plausible values are also described. These methods were very similar to the ones used in PISA-D Strand A/B (OECD, 2019; Chapter 9), PISA 2015 (OECD, 2017; Chapter 9), and PISA 2018 (OECD, 2020; Chapter 9).

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 and to ensure that the test design had been implemented as intended. The following subsections give an overview of these analyses and their results. Overall, some issues were found in the quality of the data, such as a smaller sample size than intended, uneven distribution of forms, and the miscalculation of age for a small group of respondents. Nonetheless, the data were considered to be of sufficient quantity and quality for the scaling, population modelling, and estimation of plausible levels.

Data yield

Participating countries were required to sample a minimum of 1 600 respondents between the ages of 14 and 16 who were either enrolled in school at grade 6 or below, or outside of the school system. Additionally, each country aimed to have at least 1 300 respondent who passed the core cognitive assessment. Note that literacy-related non-respondents (LRNR) were included in the analyses if they met the eligibility criteria, but those who refused to take the assessment or were unable to take it for any other reason were excluded because no information was available to assess their skills. As presented in Table 10.1, the target total sample size was not met in Honduras and Paraguay, while the target for the number of respondents that passed the core cognitive assessment was not met in Guatemala, Honduras, and Paraguay.

Table 10.1 Total sample size and number of respondents that passed the core cognitive assessment

Country	Total sample size	Pass core
Guatemala	1 749	1 208
Honduras	1 281	1 104
Panama	2 055	1 448
Paraguay	1 002	500
Senegal	2 103	1 672

In each country, the target was also to have an average of 650 responses per main assessment item (i.e. items that were presented to those who passed the core cognitive assessment), which was necessary for the stable estimation of group-specific item parameters, as explained in the section on detecting and handling misfit for groups. Table 10.2 presents the average number of responses, by country, per main assessment item in each cluster. The target number of responses per main assessment item was not met in Guatemala, Honduras, and Paraguay for all clusters.

Table 10.2 Average number of responses per item by cluster

Country	Cluster					
	Math 1	Math 2	Math 3	Reading 1	Reading 2	Reading 3
Guatemala	589	578	613	536	556	558
Honduras	541	538	528	540	547	494
Panama	659	672	695	665	650	633
Paraguay	238	225	223	220	220	206
Senegal	803	788	816	791	754	771

While analysing the data, we found that the date and time of the internal clock were incorrect for some of the tablets used in the assessment. Since the internal clock had been used to automatically calculate the respondent's age, one of the criteria used for determining eligibility for the assessment, some respondents that were eligible for the assessment may have been mistakenly excluded, while others ineligible for the assessment may have been mistakenly included. Table 10.3 presents the number and percent of the assessments with a recorded assessment date that was outside the appropriate assessment date range for each country-by-language group. This gives insights into the percent of tablets that may have had an incorrect internal clock, which may have resulted in the incorrect inclusion or exclusion of respondents.

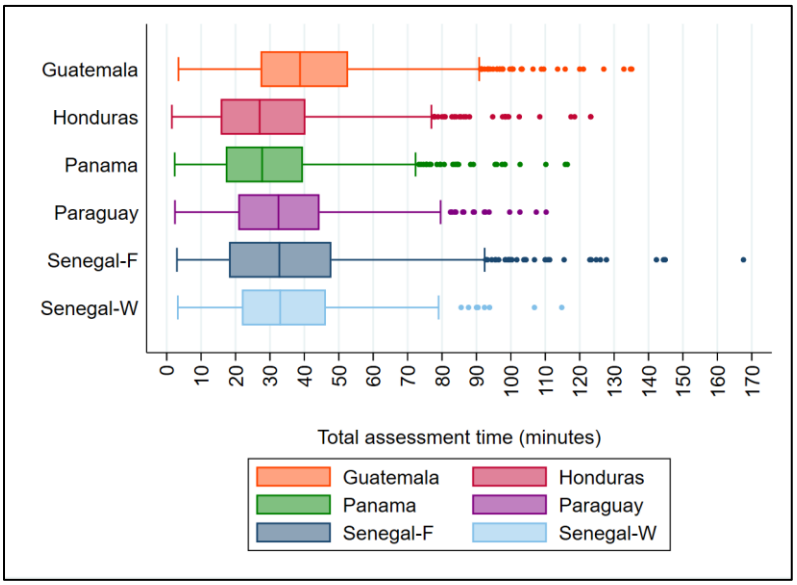
Table 10.3 Recorded assessment dates outside the appropriate assessment date range

Group	Number of assessments outside appropriate date range	Percent of assessments outside appropriate date range
Guatemala	41	2.3
Honduras	126	9.8
Panama	35	1.7
Paraguay	1	0.1
Senegal-French	148	8.7
Senegal-Wolof	42	10.7

Assessment time

For respondents who passed the core cognitive assessment, the assessment was designed to take 45 minutes (10 minutes for the core cognitive assessment and 35 minutes for the Math, Reading, and Reading Components clusters). Among the respondents who passed the core cognitive assessment, the median total assessment time was 39 minutes in Guatemala, 27 minutes in Honduras, 28 minutes in Panama, 32 minutes in Paraguay, and 33 minutes in both Senegal-French and Senegal-Wolof.[1] Figure 10.1 presents the distribution of the total assessment times for the respondents who passed the core cognitive assessment, disaggregated by country-by-language group.

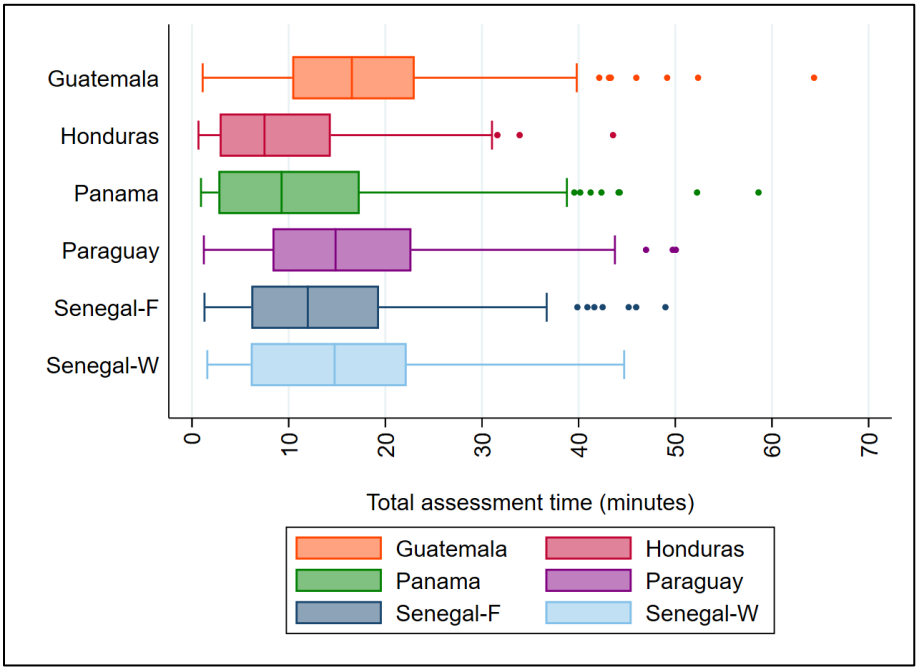
Figure 10.1 Distribution of total assessment time in minutes for respondents who passed the core cognitive assessment



[1] Please note that assessment time was computed by aggregating the time students spent on each item across visits—this total item time variable "TT" is reported in the process data public use (PUF) file.

For the respondents who failed the core cognitive assessment, the assessment was designed to take 25 minutes (10 minutes for the core cognitive assessment and 15 minutes for the Reading Components clusters). The median total assessment time for those who failed the core cognitive assessment was 17 minutes in Guatemala, 7 minutes in Honduras, 9 minutes in Panama, 15 minutes in Paraguay, 12 minutes in Senegal-French, and 15 minutes in Senegal-Wolof. Figure 10.2 presents the distribution of the total assessment times for the respondents who failed the core cognitive assessment, disaggregated by country-by-language group.

Figure 10.2 Distribution of total assessment time in minutes for respondents who failed the core cognitive assessment



Test administration

Within each country, the mechanism to randomize the assignment of forms was not fully implemented to ensure the even distribution of forms (i.e. many interviewers started using the same form in the rotation). As a result, some forms were used with more respondents than other forms, and consequently, some items were presented to more respondents than other items. This deviation from the procedures is not expected to affect the analysis of the results. Figure 10.3 shows the proportion of respondents that received each of the forms in each country-by-language group, among the respondents that passed the core cognitive assessment. If the forms had been distributed evenly, each form would have been distributed to about 8.3% of the respondents that passed the core cognitive assessment in each country-by-language group, represented by the red horizontal line.

Figure 10.3 Proportion of respondents that received forms 1 to 12 (among the respondents that passed the core cognitive assessment)

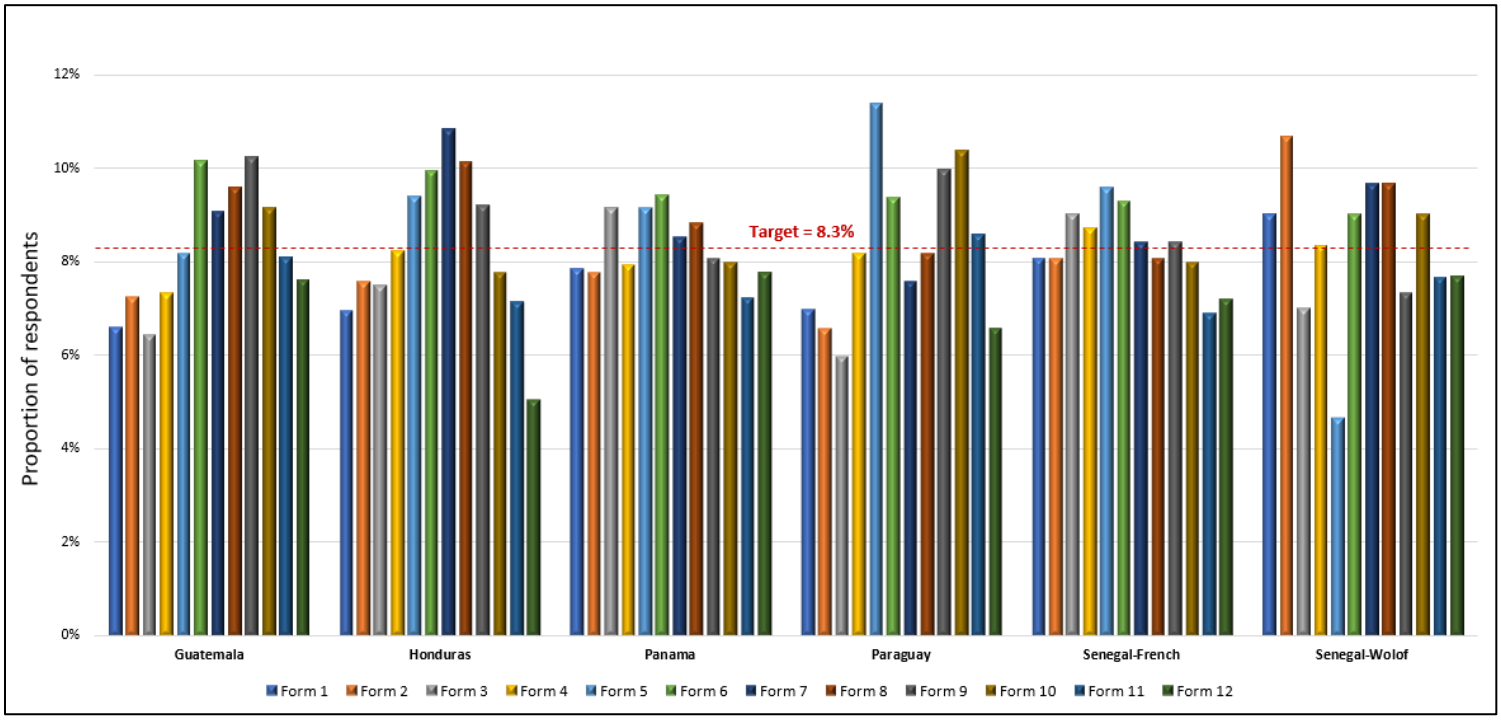
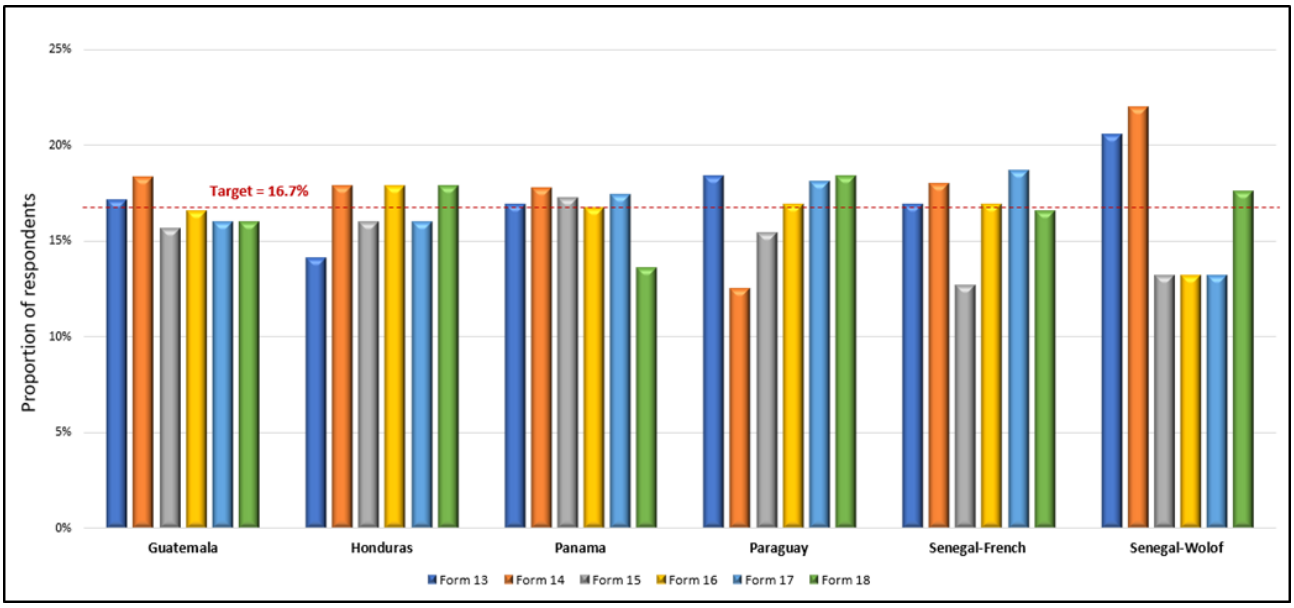


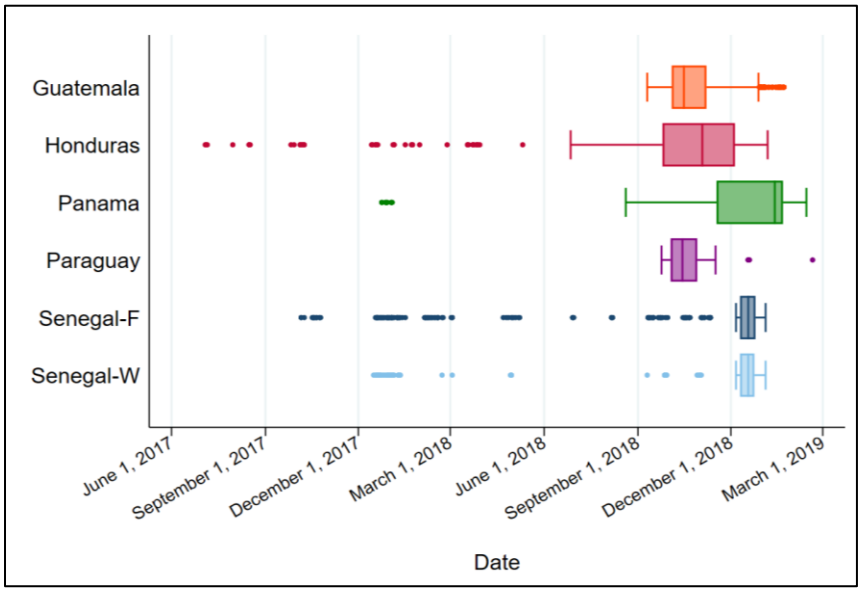
Figure 10.4 shows the proportion of respondents that received forms 13 to 18, among the respondents that failed the core cognitive assessment. If the forms had been distributed evenly, each form would have been distributed to 16.7% of the respondents that failed the core cognitive assessment in each country-by-language group, represented by the red horizontal line. This deviation from the procedures is also not expected to affect the analysis of the results.

Figure 10.4 Proportion of respondents that received forms 13 to 18 (among the respondents that failed the core cognitive assessment)



Note also that in some countries, a large proportion of the assessments were conducted during a time frame that was shorter than expected. Figure 10.5 presents a boxplot of the assessment dates for each group, with the length of the box corresponding to the time frame during which the middle 50% of the assessments were conducted – 34 days for Guatemala, 71 days for Honduras, 72 days for Panama, 26 days for Paraguay, and 15 days for both Senegal-French and Senegal-Wolof. Figure 10.5 includes cases in which the internal clock of the tablet was incorrect, which explains some of the outliers.

Figure 10.5 Distribution of assessment dates



ITEM-LEVEL AND CLUSTER-LEVEL ANALYSES

Classical test theory statistics

For each country-by-language group, analyses were conducted at the item- and cluster-level. Sampling weights were used for all analyses. These statistics were examined to identify any outlier items (i.e. items that did not function in a comparable way across the countries), machine-scoring issues, as well as other technical issues.

The item-level statistics that were examined were:

- Percent correct: For dichotomous items, the percent of respondents that were presented and responded to the item correctly. For polytomous items, the weighted percent of respondents that responded to the item correctly (with a lower weight for respondents that received partial credit for the item). Item responses classified as not-reached were excluded from the calculation of percent correct statistics.
- Percent not reached: Percent of respondents that were supposed to be presented the item according to the design, but were not, because they decided to not continue with the assessment.
- Percent omitted: Percent of respondents that were presented the item but did not provide a response.
- R-biserial: Correlation between respondents' performance on an individual item and their total score in the cluster.

When classifying item responses, an item response was considered omitted when there was no response to the item, but there was a valid response in one or more subsequent items in the assessment form. An item was considered not reached when there was no valid response for that item or for any of the subsequent items in the assessment form.

Tables 10.4 to 10.7 present the summary results for the statistics above, disaggregated by country-by-language group and cluster. The statistics were averaged across all items in the cluster, applying sample weights within each country-by-language group.

Table 10.4 Percent correct

Country	Cluster										
	Core	Math 1	Math 2	Math 3	Reading 1	Reading 2	Reading 3	RC - Sentence	RC - Passage (A)	RC - Passage (B)	RC - Passage (C)
Guatemala	36.4	13.5	14.3	11.3	23.2	25.6	19.8	66.1	65.8	55.0	60.2
Honduras	57.5	22.2	20.7	19.9	29.2	29.4	25.1	72.8	75.7	67.5	71.1
Panama	42.6	19.0	17.2	13.8	29.5	29.8	22.6	69.3	66.6	60.5	62.3
Paraguay	27.8	14.7	13.3	6.4	21.6	23.7	14.4	63.8	62.7	51.9	53.5
Senegal-F	47.9	17.3	15.0	11.8	24.1	27.1	17.0	62.1	48.7	47.2	49.2
Senegal-W	37.2	14.7	12.9	9.5	22.2	24.3	23.3	60.7	43.8	45.0	41.9

Table 10.5 Percent not reached

Country	Cluster										
	Core	Math 1	Math 2	Math 3	Reading 1	Reading 2	Reading 3	RC - Sentence	RC - Passage (A)	RC - Passage (B)	RC - Passage (C)
Guatemala	0.7	1.8	3.4	2.6	1.8	2.7	3.0	0.1	0.0	0.0	0.0
Honduras	0.4	0.8	1.4	1.1	0.7	1.0	2.5	0.1	0.0	0.0	0.0
Panama	0.6	2.2	2.2	3.0	2.4	1.8	5.6	0.0	0.0	0.0	0.0
Paraguay	0.4	4.3	3.3	5.7	2.8	1.3	7.5	0.1	0.0	0.0	0.0
Senegal-F	0.0	1.2	1.9	1.4	1.1	1.8	3.5	0.0	0.0	0.0	0.0
Senegal-W	0.1	1.4	1.9	0.1	0.7	1.7	2.1	0.0	0.0	0.0	0.0

Table 10.6 Percent omitted

Country	Cluster						
	Core	Math 1	Math 2	Math 3	Reading 1	Reading 2	Reading 3
Guatemala	13.8	18.9	23.6	23.5	12.7	18.1	24.0
Honduras	9.0	9.7	14.3	12.1	9.1	12.4	16.5
Panama	27.1	26.5	32.3	29.8	23.4	23.8	26.7
Paraguay	30.3	26.7	35.6	42.0	25.2	28.9	36.2
Senegal-F	9.7	16.3	25.6	19.3	18.0	19.9	23.1
Senegal-W	12.2	15.8	28.2	18.3	15.0	18.2	18.7

Note: Respondents did not have the option to omit items in the Reading Components cluster, so those columns are excluded from the table.

Table 10.7 R-biserial

Country	Cluster										
	Core	Math 1	Math 2	Math 3	Reading 1	Reading 2	Reading 3	RC - Sentence	RC - Passage (A)	RC - Passage (B)	RC - Passage (C)
Guatemala	0.74	0.69	0.66	0.76	0.66	0.76	0.75	0.44	0.78	0.64	0.78
Honduras	0.73	0.60	0.67	0.66	0.72	0.77	0.69	0.50	0.90	0.74	0.89
Panama	0.82	0.70	0.85	0.73	0.77	0.80	0.82	0.52	0.84	0.72	0.85
Paraguay	0.75	0.65	0.67	0.75	0.81	0.82	0.82	0.45	0.77	0.57	0.76
Senegal-F	0.73	0.57	0.74	0.71	0.71	0.75	0.72	0.43	0.73	0.60	0.66
Senegal-W	0.69	0.55	0.83	0.57	0.65	0.70	0.72	0.43	0.64	0.54	0.73

Position effects

Item position effects resulting from a cluster being placed in different positions in different test forms are a common concern in large-scale assessments because substantial position effects can increase measurement error and introduce bias. To verify that the position effects were tolerable, the extent to which the position of a cluster affected the percent correct was examined.

Table 10.8 presents the average percent correct for each domain in each cluster position. The values were averaged across all items in the domain and across all countries (with each country weighted equally and using sample weights within each country). The last column presents the difference in the percent correct between cluster position 1 and cluster position 3 for each domain. The core cognitive assessment was not included in the analysis, because it was always presented first to all respondents. For Math and Reading, the analysis only included the respondents who passed the core cognitive assessment, because those who failed the core cognitive assessment were not presented with additional items from these domains. For the Reading Components Sentence Processing cluster, the analysis only included respondents who passed the core cognitive assessment, because this cluster was presented in the same position to respondents who failed the core cognitive assessment. For the Reading Components Passage Comprehension cluster, the analysis was conducted separately for those who passed the core cognitive assessment and those who failed it.

The differences in the percent correct between cluster position 1 and cluster position 3 ranged from -2.8 percentage points (for the Reading Components Passage Comprehension cluster, among the respondents who passed the core cognitive assessment) to 3.0 percentage points (for the Reading cluster, among the respondents who passed the core cognitive assessment). Note that when calculating percent correct statistics for the purpose of calculating the position effects, not-reached items were excluded from the denominator, as described earlier in this chapter.

Note also that the average percent correct by cluster position presented in Table 10.8 is not directly comparable with the results from PISA-D Strand A/B and PISA because the definition of a respondent who did not reach an item was not the same across these three assessments. In PISA-D Strand A/B and PISA, respondents who did not respond to an item as well as any subsequent items in the *cluster* were counted as respondents who did not reach the item, while in PISA-D Strand C, those who did not respond to an item as well as any subsequent items in the *form* were counted as respondents who did not reach the item. In addition, other differences between the three assessments may have affected the number of respondents that did not reach an item: PISA-D Stand C only had 3 clusters, while PISA-D Strand A/B and PISA had 4 clusters; PISA-D Strand C had no break in the middle of the assessment, while PISA-D Strand A/B and PISA had a break after the second cluster. Since the number of respondents who did not reach an item are excluded from the denominator when calculating the percent correct for an item, these differences make it inappropriate to compare the cluster-level percent correct in PISA Strand C to the results from PISA-D Strand A/B and PISA.

Table 10.8 Average percent correct by cluster position

Domain	Position 1	Position 2	Position 3	Position 3 – Position 1 (percentage points)
Math (Pass core)	16.1%	14.6%	15.1%	-1.0
Reading (Pass core)	22.6%	23.6%	25.6%	3.0
RC - Sentence Processing (Pass core)	70.3%	69.9%	68.6%	-1.7
RC - Passage Comprehension (Pass core)	70.2%	70.6%	67.4%	-2.8
RC - Passage Comprehension (Fail core)	46.7%	47.1%	46.7%	0.0

IRT MODEL FOR SCALING

The test design for PISA-D Strand C was based on a combination of adaptive testing (through the core module) and a variant of matrix sampling where each respondent was administered a subset of items from the total item pool. That is, different respondents answered different yet overlapping sets of items. This design was necessary to represent the broad measurement constructs with many more items than an individual respondent was able to respond to in a testing session, as well as to adapt the test difficulty to the target sample of PISA-D Strand C. While this design has its advantages, a common disadvantage is that it makes it inappropriate to use any statistic based on the total number of correct responses. Differences in total scores, or statistics based on them, among respondents who took different sets of items may be due to variations in the difficulty of the test forms.

The limitations of scoring methods based on the number or percent of correct responses can be overcome by using IRT scaling. When responding to a set of items requires a given skill, the response patterns should show regularities that can be modelled using the underlying commonalities (i.e. a latent trait called θ) among the items. This regularity can be used to characterise respondents as well as items in reference to a common scale, even if all respondents do not take identical sets of items. It also makes it possible to describe the distribution of

performance in a population or subpopulation and to estimate the relationships between proficiency and background variables as accurately as possible.

The calibration and scaling methods used for PISA-D Strand C followed the approach used for PISA-D Strand A/B, PISA 2015, and PISA 2018. Specifically, the two-parameter logistic model (2PLM; Lord & Novick, 1968), which is an IRT model that is a generalisation of the Rasch model, was used to scale dichotomously scored items. Similar to the Rasch model, the 2PLM assumes that the probability of response *x* to an item *i* by a respondent *s* depends on the difference between the respondent's proficiency (θ_s) and the difficulty of the item (β_i). In addition, for every item, the 2PLM allows the association between this difference and the response probability to depend on an additional item discrimination parameter (α_i), characterising the sensitivity of the item to proficiency. Thus, in the 2PLM, the response probability to an item is a function of a person parameter and two item parameters, as expressed in the following formula:

$$P(x_{is} = 1|\theta_s, \beta_i, \alpha_i) = \frac{\exp(\mathrm{D}\alpha_i(\theta_s - \beta_i))}{1 + \exp(\mathrm{D}\alpha_i(\theta_s - \beta_i))} \qquad (10.1)$$

Note that D is a constant of arbitrary size, often 1.0 or 1.7, depending on the parameterisation used in the software. In the case of PISA-D Strand C, a value of 1.7 was used for D, following the parameterisation used in PISA 2015 (OECD, 2017) and PISA 2018 (OECD, 2020). For α_i > 0, the function is a monotone increasing function with respect to θ. In other words, the conditional probability of a correct response increases as the value of θ increases. One important special case of the model is when α_i = 1 for all items, in which case the model is equivalent to a Rasch model. Thus, the Rasch model is a special case of the 2PLM, and the two models differ only when the optimal estimates for the slope parameter (α_i) are different across the items.

For polytomously scored items (i.e. items with more than two ordered response categories), the generalised partial credit model (GPCM; Muraki, 1992) was used for scaling. The GPCM is similar to the 2PLM, but it can be used to scale both dichotomously scored items as well as polytomously scored items. (Note that when the GPCM is used to scale dichotomously scored items, it is equivalent to the 2PLM.) For an item *i* with m_i+1 ordered categories, the GPCM can be written as:

$$P(x_{is} = k|\theta_s, \beta_i, \alpha_i, d_i) = \frac{\exp\{\sum_{r=0}^{k} \mathrm{D}\alpha_i\,(\theta_s - \beta_i + \boldsymbol{d}_{ir})\}}{\sum_{u=0}^{m_i} \exp\{\sum_{r=0}^{u} \mathrm{D}\alpha_i\,(\theta_s - \beta_i + \boldsymbol{d}_{ir})\}} \qquad (10.2)$$

where $\boldsymbol{d}_i$ is a vector of category threshold parameters.

A central assumption of most IRT models is conditional independence, sometimes referred to as local independence. Under this assumption, item response probabilities depend only on θ and the specified item parameters—there is no dependence on any demographic characteristics of the respondents, responses to any other items presented in the assessment, or the survey administration conditions. Another important assumption is that the primary (often single) score for each domain measured can be accounted for by a dominant latent variable, θ. In other words,

it is assumed that the scale is unidimensional. When these assumptions are satisfied, the joint probability of a particular response pattern $\boldsymbol{x}$=(x_1,…,x_n) across a set of n items can be expressed as:

$$P(\boldsymbol{x}|\theta,\boldsymbol{\beta},\boldsymbol{\alpha}) = \prod_{i=1}^{n} P_i(\theta)^{x_i}\,(1-P_i(\theta))^{1-x_i} \qquad (10.3)$$

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 respondents provide their answers independently of one another and that the respondents' proficiencies are sampled from a distribution, (θ). The likelihood function, therefore, is characterised as:

$$P(\boldsymbol{X}|\boldsymbol{\beta},\boldsymbol{\alpha}) = \prod_{j=1}^{J} \int \left(\prod_{i=1}^{n} P_i(\theta)^{x_{ij}}\,(1-P_i(\theta))^{1-x_{ij}} \right) f(\theta)\,d\theta \qquad (10.4)$$

Given a dataset of scored item responses and a choice of item response models (i.e. the 2PLM or the GPCM models described above), the item parameters and the person latent traits can be estimated by maximising this function.

The scaling of PISA-D Strand C was carried out separately for Reading (including Reading Components) and Math using the software *mdltm* (von Davier, 2005) which provides marginal maximum likelihood (MML) estimates obtained using customary expectation-maximisation (EM) methods with optional acceleration. Regarding the treatment of missing responses, any missing response prior to a valid response in a form was defined as an omitted response and treated as an incorrect response. In contrast, sequential missing responses at the end of each form (regardless of the domain) were treated as not reached or not administered (i.e. these items were treated as missing), so they had no impact on the IRT scaling.

Developing a common scale between PISA-D Strand C and PISA

The primary goal of scaling PISA-D Strand C was to provide a reliable and valid link to the PISA scale (i.e. the scores from both assessments can be located on a comparable scale) by linking PISA-D Strand C to PISA-D Strand A/B through fixed item parameter linking. Therefore, all of the items in PISA-D Strand C, with the exception of one item, were selected from PISA-D Strand A/B and the scoring rules were also the same as those applied in PISA-D Strand A/B.[2] The one new item in PISA-D Strand C was an item in the Reading Components Sentence Processing cluster which had been developed for PISA-D Strand A/B but was eventually dropped because of a negative slope parameter when scaled with PISA-D Strand A/B data. Note that PISA-D Strand A/B

[2] Note that in PISA-D Strand A/B, the item parameters for items that had not been sourced from PISA were estimated with respondents only from Spanish-speaking countries. Also, PISA-D Strand A/B was administered as a paper-based assessment (PBA) instead of using tablets.

had already been linked to PISA 2015 and comprised items from PISA 2015, PISA for Schools, the Programme for the International Assessment of Adult Competencies (PIAAC), the STEP Skills Measurement Program, and the Literacy Assessment and Monitoring Programme (LAMP; OECD, 2019).[3]

Tables 10.9 to Table 10.11 provide summary information on the original source of the Math, Reading, and Reading Components items, respectively. Details on the source of each item as well as the international item parameters are presented in Annex A of this report. All of the items, except for the one new Reading Components item, served as linking items between PISA-D Strand C and PISA-D Strand A/B. In addition, over 40 percent of the items in Math and over 60 percent of the items in Reading also served as linking items between PISA-D Strand C and PISA 2015, thereby creating comparable scales between PISA-D Strand C, PISA-D Strand A/B, and PISA.

Table 10.9 Source of Math items

Source	Freq.	Percent
PISA-D Strand A/B & PISA 2015 Trend (fixed to PISA 2015 parameters)	15	43
PISA-D Strand A/B & PISA 2015 Trend (newly estimated parameters)	3	9
PISA-D Strand A/B & PISA for Schools	6	17
PISA-D Strand A/B & PIAAC	11	31
Total	**35**	**100**

Table 10.10 Source of Reading items

Source	Freq.	Percent
PISA-D Strand A/B & PISA 2015 Trend (fixed to PISA 2015 parameters)	14	64
PISA-D Strand A/B & PISA for Schools	2	9
PISA-D Strand A/B & PIAAC	2	9
PISA-D Strand A/B & LAMP	4	18
Total	**22**	**100**

Table 10.11 Source of Reading Components items

Source	Freq.	Percent
PISA-D Strand A/B	49	98
New item	1	2
Total	**50**	**100**

For the IRT scaling, at the beginning of the scaling process, all the item parameters were fixed to the parameters that had been obtained in PISA-D Strand A/B (including the group-specific item parameters). The next section describes how group-specific misfit was handled when the PISA-D

[3] In PISA-D Strand A/B, item parameters for approximately 40% of the Math items and 60% of the Reading items were identical to the parameters that had been estimated in PISA 2015, providing the strongest link among the three different assessments (i.e. PISA 2015, PISA-D Strand A/B, and PISA-D Strand C).

Strand A/B parameters did not fit the data from PISA-D Strand C. Multiple group concurrent calibration was also used, with each country-by-language group defined as a separate group. Note that for the purposes of scaling, Senegal was divided into the French-speaking and Wolof-speaking group, while all the other countries that participated in PISA-D Strand C only included a single language version of the group. For the one new item in the Reading Components Sentence Processing cluster, equality constraints were imposed so that common item parameters would be estimated for all the country-by-language groups.

Detecting and handling item misfit for certain groups

As mentioned above, at the beginning of the scaling process, all the item parameters (including the group-specific item parameters) were fixed to the parameters that had been obtained in PISA-D Strand A/B. Subsequently, to examine how well the PISA-D Strand A/B item parameters fit the data for PISA-D Strand C, the root mean square deviation (RMSD) was examined for each item-by-group combination. The RMSD quantifies the absolute difference between the model-based item characteristic curve (based on parameters obtained from PISA-D Strand A/B) and the empirical item characteristic curve (from PISA-D Strand C) and is calculated using the following formula for each item:

$$RMSD_g = \sqrt{\int \left[p_g^{obs}(\theta) - p_g^{exp}(\theta)\right]^2 f_g(\theta) d\theta} \qquad (10.5)$$

where g = 1, … , G is a country-by-language group; $p_g^{obs}(\theta)$ and $p_g^{exp}(\theta)$ are the observed and expected probability of a correct response, respectively, given the proficiency θ; and $f_g(\theta)$ is the group-specific density distribution of the group members' ability scale. The values of RMSD range from 0 to 1, with a higher value indicating a higher level of misfit for the item-by-group combination.

Similar to the approach used in PISA 2018, when the RMSD for an item-by-group combination was over 0.4 or when the slope parameter was close to 0, the item was excluded from the scaling for the group. Additionally, if the RMSD for an item-by-group combination was over 0.15 but less than or equal to 0.4, unique item parameters were estimated for the group until all item-by-group combinations had an RMSD of 0.15 or below. Also, if more than one group exhibited misfit for an item and the groups had a similar level and direction of misfit, the item parameters for the groups were estimated together. The process of detecting misfit and assigning new item parameters to the groups exhibiting misfit was carried out using an automatic algorithm in *mdltm*. It should be noted that for some items, the power to detect group-specific misfit may have been lower than optimal due to the reduced sample size noted earlier.

The scaling methods described above are based on models originally developed within the framework of IRT that have evolved into very flexible approaches for the analysis of large-scale multilevel categorical data (e.g. Skrondal & Rabe-Hesketh, 2004; von Davier & Yamamoto, 2004, 2007; Adams, Wu, & Carstensen, 2007). Extensive descriptions of the scaling methods and procedure used in PISA-D Strand C are provided in Yamamoto and Mazzeo (1992); Mislevy and

Sheehan (1987); Glas and Verhelst (1995); and Adams, Wilson, and Wu (1997). More recent overviews of the different aspects of the methodology can be found in von Davier, Sinharay, Oranje, and Beaton (2006); Glas and Jehangir (2014); Weeks, von Davier, and Yamamoto (2014); von Davier and Sinharay (2014); and Mazzeo and von Davier (2014). In order to account for cultural and language differences in the multiple populations tested, procedures outlined in Glas and Verhelst (1995), Yamamoto (1997), Glas and Jehangir (2014), and Oliveri and von Davier (2011, 2014) were applied.

POPULATION MODELLING

In international large-scale assessments such as PISA-D Strand C, test forms are kept relatively short to minimise individuals' response burden. As a consequence, only a subset of the items is administered to each respondent, resulting in a high level of missing data at the individual level. To account for this uncertainty at the individual level and to increase the accuracy of the estimates of the multivariate proficiency distributions for the population, the plausible values methodology was used.

Plausible values

Plausible values were generated through population modelling which is a combination of an IRT model and a multivariate latent regression model, taking into account respondents' responses to the background questionnaires (BQ) and cognitive items, as well as the covariance between the assessed domains. Specifically, in PISA-D Strand C, data collected in the BQ were contrast coded (refer to Annex B for the contrast coding used in the conditioning model), then transformed by means of a principal component analysis (PCA) in order to reduce the number of variables used in the analysis. Once the principal components were calculated, a certain number was used in the multivariate latent regression model as predictors. The number of principal components used was that which explained 80% of the variance of the BQ variables or was equal to 5% of the respondent sample size, whichever corresponded to a smaller number of principal components. Subsequently, the latent regression parameters Γ (regression coefficients) and Σ (residual variance-covariance matrix) were estimated. In this process of estimating the regression coefficients, only respondents who attempted at least five items in Math or at least five items in Reading were included in the analysis, and the item parameters were fixed to the values that had been obtained through IRT scaling in the previous item calibration stage. Lastly, for each respondent (including the respondents who were not included in estimation of the latent regression parameters), a plausible value was drawn from the normal approximation of the posterior distribution of the proficiency variable (θ), creating a set of plausible values for all respondents. This process, which was implemented using the software DGROUP (Rogers, Tang, Lin, & Kandathil, 2006), was repeated 10 times to create 10 sets of independently drawn plausible values. While each set of plausible values is equally well designed to estimate population statistics, such as group means and standard deviations, multiple plausible values are required to appropriately represent the uncertainty in the domain measured (von Davier, Gonzalez, & Mislevy, 2009).

Note that population modelling was carried out separately for each country in order to allow for between-country differences in the associations between the background variables and cognitive skills. Also, as with all international large-scale assessments, statistics based on plausible values are only intended to be used to report at the population or subpopulation levels and should never be used to draw inferences at the individual level.

For more details on the rationale behind the plausible value methodology and the computational procedure, please refer to the technical reports for PISA-D Strand A/B (OECD, 2019), PISA 2015 (OECD, 2017), and PISA 2018 (OECD, 2020), as well as papers by Mislevy (1985, 1991); Mislevy and Sheehan (1987); Thomas (2002); von Davier, Sinharay, Oranje, and Beaton (2006); von Davier, Gonzalez, and Mislevy (2009); Rutkowski, Gonzalez, Joncas, & von Davier (2010); and von Davier and Sinharay (2014).

Transforming the plausible values to the PISA scale

As stated above, the goal of PISA-D Strand C was to link the results to those of PISA through PISA-D Strand A/B. The three assessments also have a common set of items with identical item parameters, and both used the same population modelling methods to generate the plausible values. For these reasons, PISA-D Strand C used the same transformation coefficients as PISA to transform the plausible values to the PISA reporting scale. Table 10.12 provides these transformation coefficients. Coefficient A adjusts for the variability (standard deviation) of the plausible values, while coefficient B adjusts for the scale location (mean).

Table 10.12 PISA transformation coefficients

Domain	A	B
Math	135.90	514.18
Reading	131.58	437.96

PLAUSIBLE LEVELS

Due to the data quality issues mentioned above, as well as the relatively high level of measurement uncertainty in PISA-D Strand C, all plausible values were categorized into plausible levels using the cut points presented in Table 10.13 and Table 10.14. In the public use file (PUF), ten plausible levels (instead of plausible values) are reported for each respondent, with the levels coded as 0 = Below Level 1c, 1 = Level 1c, 2 = Level 1b, 3 = Level 1a, and 4 = Level 2 and above. Note that respondents who did not respond to any of the cognitive items as well as the LRNR cases were automatically assigned the lowest plausible level (i.e. below level 1c) for all 10 plausible levels for both Math and Reading, because it was considered that the LRNR cases did not have the language skills required to be functional in the population.

Table 10.13 Cut points for each proficiency level for Math

Level	Cut points on the PISA scale
Level 2 and above	420.07 or higher
Level 1a	357.77 or higher
Level 1b	295.47 or higher
Level 1c	233.17 or higher
Below level 1c	Below 233.17

Table 10.14 Cut points for each proficiency level for Reading

Level	Cut points on the PISA scale
Level 2 and above	407.47 or higher
Level 1a	334.75 or higher
Level 1b	262.04 or higher
Level 1c	189.33 or higher
Below level 1c	Below 189.33

ANALYSIS OF DATA WITH PLAUSIBLE LEVELS

Working with plausible levels requires a similar approach as working with plausible values, that is, these should not be understood as individual-level categorizations that tell us with any certainty the level of an individual's performance. When working with plausible levels, the analysis should be conducted separately with each plausible level classification, then the 10 results should be averaged. The imputation variance is calculated by multiplying the variance of the 10 results by an expansion factor, in this case, equal to (1 + 1/M), where M is the number of plausible values. Chapter 12 of this technical report elaborates on how to work with plausible levels, calculate the imputation variance, and obtain the error of a statistic by combining the imputation variance with the sampling variance.

REFERENCES

Adams, R. J., Wilson, M. R., & Wu, M. L. (1997). Multilevel item response models: An approach to errors in variables regression. Journal of Educational and Behavioural Statistics, 22, 46–75.

Adams, R. J., Wu, M. L., & Carstensen, C. H. (2007). Application of multivariate Rasch models in international large-scale educational assessments. In M. von Davier, & C. H. Carstensen (Eds.), Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 271-280). New York, NY: Springer.

Glas, C. A. W., & Jehangir, K. (2014) Modeling country specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of international large-scale assessment (pp. 97-115). Boca Raton, FL: 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). New York, NY: Springer.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mazzeo, J., & von Davier, M. (2014). Linking scales 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. 229-258). Boca Raton, FL: CRC Press.

Mislevy, R. J. (1985). Estimation of latent group effects. Journal of the American Statistical Association, 80(392), 993–997.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. Psychometrika, 56(2), 177–196.

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 (Report No. 15-TR-20). Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16(2), 159–177.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. Psychological Test and Assessment Modeling, 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. doi:10.1080/15305058.2013.825265

Organization for Economic Co-operation and Development (OECD). (2017). PISA 2015 technical report. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2019). PISA for Development technical report. Paris, France: OECD. Retrieved from https://www.oecd.org/pisa/pisa-for-development/pisafordevelopment2018technicalreport/

Organization for Economic Co-operation and Development (OECD). (2020). PISA 2018 technical report. Paris, France: OECD.

Rogers, A., Tang, C., Lin, M. J., & Kandathil, M. (2006). DGROUP [Computer software]. Princeton, NJ: Educational Testing Service.

Rutkowski, L., Gonzalez, E., Joncas, M. & von Davier, M. (2010) International large-scale assessment data: Issues in secondary analysis and reporting. Educational Researcher, 39(2), 142-151.

Skrondal, A. & Rabe-Hesketh, S. (2004). Generalized latent variable modeling: Multilevel, longitudinal and structural equation models. Boca Raton, FL: Chapman & Hall/CRC.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. Psychometrika, 67(1), 33-48.

von Davier, M. (2005). mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: 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.

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). Boca Raton, FL: CRC Press.

von Davier, M. Sinharay, S., Oranje, A., & Beaton, A. (2006) Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao and S. Sinharay (Eds.), Handbook of statistics: Vol. 26, Psychometrics (pp. 1039-1055). Amsterdam, Netherlands: Elsevier.

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.

von Davier, M. & Yamamoto, K. (2007). Mixture distribution Rasch models and Hybrid Rasch models. In M. von Davier & C. H. Carstensen (Eds.), Multivariate and mixture distribution Rasch models (pp. 99-115). New York, NY: Springer.

Weeks, J., von Davier, M., & Yamamoto, K., (2014). Design considerations for the Program for International Student Assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 259-275). Boca Raton, FL: CRC Press.

Yamamoto, K. (1997). A chapter: Scaling and scale linking. International Adult Literacy Survey (IALS) technical report. Ottawa, Canada: Statistics Canada.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. Journal of Educational Statistics, 17(2), 155-174.