Chapter 9: SCALING PISA-D COGNITIVE DATA

INTRODUCTION

The test design for PISA-D is similar to the one used for PISA. It is based on a variant of matrix sampling where each student is administered a relatively small subset of items from the total item pool. That is, different students answer different yet overlapping sets of items and equivalent amounts of data is collected on each item. This design is necessary to represent the broad measurement constructs with many more items than an individual student would be able to respond to in a testing session.

But this design also makes it inappropriate to use any statistic based on the number of correct responses in reporting the survey results. Differences in total scores, or statistics based on them, among students who took different sets of items may be due to variations in difficulty of the test forms. Unless one makes very strong assumptions—for example, that the different test forms are perfectly parallel—the performance of two groups assessed in a matrix sampling arrangement cannot be directly compared using total-score statistics. Moreover, item-by-item reporting ignores the dissimilarities of proficiencies of subgroups to which the set of items was administered. Finally, using the average percentage of items answered correctly to estimate the mean proficiency of students in a given subpopulation does not provide any other information about the distribution of skills within that subpopulation (e.g., variances).

The limitations of number or percent correct scoring methods can be overcome by using item response theory (IRT) scaling. When responding to a set of items requires a given skill, the response patterns should show regularities that can be modeled using the underlying commonalities among the items. This regularity can be used to characterise students as well as items in reference to a common scale, even if all students do not take identical sets of items. It also makes it possible to describe distributions of performance in a population or subpopulation and to estimate the relationships between proficiency and background variables as accurately as possible.

In the following sections, a description of the analyses and the results obtained in PISA-D are provided. The methods and analyses used closely followed the ones used in PISA. Overviews of the data yield and data quality analyses, classical item analyses, IRT scaling, and population modeling methods used to produce plausible values needed for secondary analyses are also provided.

DATA YIELD AND DATA QUALITY

Before data were used for scaling and population modeling, different analyses were carried out to examine the quality of data and to ensure that data met the test design criteria. The following subsections give an overview of these analyses and their results. Overall, the data quality was confirmed, and data was approved for scaling.

Targeted sample size and data yield

Targeted sample size

The Main Survey assessment design for PISA-D covered the cognitive domains of Reading,[1] Mathematics, and Science, with equal weights for each of the three domains (i.e., no major/minor domain distinction). Participating countries were required to sample a minimum of 150 schools with a target of 35 students per school for a total sample of approximately 5,250 students who were age 15. PISA-D was administered as a paper-based assessment (PBA) and was designed with the total testing time for measuring the three domains to be two hours for each student. Students responded to one out of 12 possible testing booklets, which consisted of four 30-minute clusters assembled from two of the three core domains, resulting in one hour of assessment time per domain, with a total of two hours of testing time per student. For the booklets containing the Reading domain, one Reading Components cluster was always placed in front of the Reading cluster. Within these clusters, a majority of items were selected from previous cycles of PISA but were complemented with existing materials from other surveys, including PISA for Schools, PIAAC (Programme for the International Assessment of Adult Competencies), the STEP Skills Measurement Program, and LAMP (Literacy Assessment and Monitoring Programme).

Data yield

Table 9.1 below shows the sample sizes and assessment languages for all seven participating countries. Note that a student was considered a "respondent" and was included in the analysis if the student responded to at least half of the cognitive items in his or her particular booklet in any domain. When less than half of the cognitive items were answered, the student had to respond to at least one cognitive item and possess data from the context questionnaire session. Due to population size and operational issues, not all countries reached the desired sample size.

Table 9.1 **Language, number of schools, sample size per country**

Country	Language	N of Schools	N total
Cambodia	Khmer	170	5162
Ecuador	Spanish	173	5664
Guatemala	Spanish	191	5100
Honduras	Spanish	213	4773
Paraguay	Spanish	205	4510
Senegal	French	162	5193
Zambia	English	186	4213

Table 9.2 presents the distribution of students per school. All countries met the technical standards of a minimum number of schools (150), weighted school response rate (85%), and weighted student response rate (80%).

[1] Reading Components are included in this scale.

Table 9.2 **Distribution of students per school**

Country	Number of Schools	Minimum # of students	Mean # of students	Median # of students	Maximum # of students	Number of schools with # of students < 15
Ecuador	173	1	32.7	38	42	25
Guatemala	191	2	26.7	30	42	40
Honduras	213	1	22.4	25	40	67
Paraguay	205	2	22	20	42	76
Cambodia	170	2	30.4	36	41	28
Senegal	162	1	32.1	36	40	17
Zambia	186	1	22.7	24	40	51

ITEM ANALYSIS

Classical test theory statistics

Classical item analyses were conducted on the items at both the national and international levels to identify any items performing as outliers, identify human- or machine-scoring issues, and identify other technical issues. Item and cluster level statistics based on observed and missing responses were provided and compared across cluster positions. These statistics were shared with countries and the OECD.

The following statistics were computed:

- item difficulties (proportion of correct responses, or P+)
- frequencies of scores (number of students attempted, correct and incorrect responses, omitted items, not-reached items)
- cluster scores (i.e., the total score within a cluster) of students with specified response types for a given item
- point biserial correlations

Statistics were compiled and examined at the aggregate level across countries and individually by country in order to identify outliers (single items that seemed to work differently across countries). Irregular cases, such as outliers or items functioning unexpectedly or with obvious scoring rule deviations, were examined. Proportion correct and missing rates of trend items were compared to results from prior PISA cycles whenever relevant.

Table 9.3 **Example output for examining response distributions**

If the highest total score is used as th											
	7	NOT RCH		OMIT	1	2	3 *	4	TOTAL	R BIS =	0.4270
ITEM 5R	N	28		34	25	298	1402	41	1800	PT BIS =	0.3056
	PERCENT	1.53		1.89	1.39	16.56	77.89	2.28	100.00	P+ =	0.7789
PM800Q01	MEAN SCORE	0.64		2.44	4.44	5.80	8.12	4.15	7.49	DELTA =	9.93
	STD. DEV.	0.77		2.19	3.86	3.49	3.76	2.87	3.89		
MC	RESP WT	0.00		0.00	0.00	0.00	1.00	0.00		ITEM WT =	0.00
If the highest total score is used as th											
	8	NOT RCH	OFF TSK	OMIT	0	1			TOTAL	R BIS =	0.4270
ITEM 5	N	28	4	30	364	1402			1800	PT BIS =	0.3056
	PERCENT	1.53	0.22	1.67	20.22	77.89			100.00	P+ =	0.7789
PM800Q01S	MEAN SCORE	0.64	5.25	2.07	5.52	8.12			7.49	DELTA =	9.93
	STD. DEV.	0.77	2.86	1.77	3.51	3.76			3.89		
MC	RESP WT	0.00	0.00	0.00	0.00	1.00				ITEM WT =	1.00

Table 9.3 is an example of classical response analysis outputs for an item. The first column identifies the item and includes its number within the cluster of items analysed together, the item code, and its type (MC for multiple choice). The part of the output identified as ITEM 5R (fifth item in the block of items analysed together; R for raw score or unscored item) provides statistics for each response option with an asterisk (*) indicating which response is correct (here, option 3) as well as the classical item statistics. The part of the output identified as ITEM 5 provides statistics for incorrect and correct answers (scored item) as well as the classical item statistics.

Row identifiers in the second column indicate the type of statistic:

1. N = number of responses for the given type, excluding those not reached
2. Percent = percent of responses for the given type
3. Mean Score = mean number correct score of the cluster for the given type
4. Std. Dev. = standard deviation of the number correct score of the cluster for the given type
5. RESP WT = response weight for the given type

The response types are:

1. NOT RCH (not reached) = students did not answer the given item or the subsequent items within that cluster
2. OFF TSK (off task) = students did not answer the question in the expected manner
3. OMIT (omit) = students did not answer the given question but answered at least one subsequent question
4. 0 = incorrect responses
5. 1 = correct responses

The values in the TOTAL column (third to the last column) are based on all categories except "NOT RCH". For example, for Item 5 (PM800Q01S), Total is the sum of OFF TSK, OMIT, 0 (Wrong) and 1 (Correct), that is, 1800 = 4 + 30 + 364 + 1402, which does not include NOT RCH, whose value is 28.

The statistics shown in the last two columns of Table 9.3 are the following:

1. R-biserial (R BIS) and R-polyserial (R POLY): R BIS is used for dichotomous items and is a statistic used to describe the relationship between performance on a single test item and a continuous criterion variable (total score on the cluster). It is an estimate of the correlation between the criterion cluster score and an unobserved normally distributed variable assumed to determine performance on the observed categorical item score. R POLY is used for polytomous items and is a generalisation of the biserial correlation for use with either dichotomous or polytomous items. This is the generalised form of the correlation with the criterion and the item score, where the item score is either (0, 1) or (0, 1, 2, 3….n) and the criterion is a continuous variable (total score on the cluster).
2. Point biserial (PT BIS) and point-polyserial (PT POLY): PT BIS is used for dichotomous items and is the Pearson product moment correlation coefficient between the dichotomous item score and the total cluster score. For polytomous items, PT POLY is used.
3. P+: This is the usual percent correct for a given item.
4. Delta: This statistic is an index of item difficulty associated with the percent correct (P+). The P+ values are converted to z-scores and are then linearly transformed to an expected value of 13.0 and a standard deviation of 4.0. Deltas ordinarily range from 6.0 for a very easy item (approximately 95% correct) to 20.0 for a very hard item (approximately 5% correct), with 13.0 corresponding to 50% correct.
5. ITEM WT: This value is the number of score levels for a scored item. For a raw item, the item weight is assigned a value of 0 to prevent the item from being double-counted in the overall cluster statistics.

Table 9.4 provides an example of the breakdown of item score categories and biserial correlations by category as well as a summary of items that were flagged for surpassing certain thresholds (the thresholds are shown in Table 9.5). In this example, the last item in the image is flagged for having an omit rate of greater than 10%, which prompted further review.

Table 9.4 **Example table providing summary item statistics**

BLOCK M1 (UNWEIGHTED)									
Item Score Category Analysis (Partial credit model)									
	Category	N	Pct. At	ct. Below	Mean	Std. Dev.	Biserial	B *	
ITEM 4	0	475	26.33	0.00	4.09	2.63			
PM5188Q01A	1	621	34.42	26.33	6.84	2.91	0.5530	-0.7708	
	2	708	39.25	60.75	10.30	3.26	0.6098	0.1759	
ITEM 5R	0	398	22.11	0.00	5.25	3.52			
PM800Q01	1	1402	77.89	22.11	8.12	3.76	0.4270	-1.2621	
ITEM 5	0	398	22.11	0.00	5.25	3.52			
PM800Q01S	1	1402	77.89	22.11	8.12	3.76	0.4270	-1.7989	
ITEM 6	0	903	50.53	0.00	4.99	2.58			
PM604P505A	1	166	9.29	50.53	8.36	3.06	0.6328	1.0841	
	2	718	40.18	59.82	10.52	3.12	0.3789	-1.6710	
BLOCK M1 (UNWEIGHTED)									
Item Analysis Flag Summary									
Item ID	Num Resp	Type	R-BIS	P-PLUS	% NOTRCH	% OFFTSK	% OMIT	% MISS	Flags
PM5188Q0	3	ECR	0.7216	0.5646	1.31	0.00	7.15	8.37	.....
PM800Q01	4	MC	0.4270	0.7789	1.53		1.89	3.39	.....
PM800Q01	2	SCR	0.4270	0.7789	1.53	0.22	1.67	3.39	.....
PM604P50	3	ECR	0.8146	0.4482	2.24	0.00	13.88	15.81	...O..

Table 9.5 **Flagging criteria for items in the item analyses**

	Criteria for flagging items
rbis/rpoly	< 0.3
P+	0.20 > P+ > .90
Omit	>10%
Off task	>10%
Not-Reached	>10%

The delta statistic, polyserial correlation, and B* are part of the standard output from the software used for the classical item analysis; however, they may not be as familiar as other statistics such as P+, R-Bis, percent not reached, and percent of omitted responses. Therefore, countries were advised to use the latter statistics when evaluating the quality of items for their sample.

Position effects

Item position effects due to each cluster appearing in different positions across test forms are a common issue of concern in large-scale assessment programs because substantial position effects can increase measurement error and introduce bias. The PISA-D Main Survey design balances

cluster position in order to control for the potential impact of item position on scores. Nevertheless, it is important to monitor the extent to which position effects impact various item statistics to ensure that these effects are tolerable. We examined the overall cluster position effects (weighting individual countries equally) in terms of: 1) proportion of correct responses by cluster (average P+), and 2) rate of omitted responses by cluster (omission rate). Note that "not reached" responses were excluded in calculating the proportion of correct responses and omitted rates.

Table 9.6 presents position effects via proportion correct by cluster for PISA-D along with relevant previous assessments. In order to establish a reference point for examining the magnitude of position effects, average P+ values were computed at the cluster level using PBA data from both PISA 2009 and 2012, and CBA (computer-based assessment) data from PISA 2015. For two PISA PBA cycles, there was an average decrease of 0.04 to 0.08 points in the average P+ metric between cluster positions 1 and 4, depending on the domain. For the PISA 2015 Main Survey CBA, the average decrease was only about 0.02 to 0.06 points in P+ values between cluster positions 1 and 4, depending on the domain. Position effects in PISA-D were slightly larger than those in PISA 2015, ranging between 0.03 and 0.08, but are similar to that of other PBA cycles.

Table 9.6 **PISA 2009, 2012 PBA and 2015 CBA average proportion correct across clusters and across countries**

	Domain	Position 1	Position 2	Position 3	Position 4	Position 4-Position 1
2009 PBA	Mathematics	0.411	0.402	0.385	0.371	-0.040
	Reading	0.584	0.559	0.534	0.501	-0.083
	Science	0.490	0.478	0.457	0.435	-0.055
2012 PBA	Mathematics	0.443	0.435	0.413	0.397	-0.046
	Reading	0.595	0.561	0.551	0.512	-0.083
	Science	0.526	0.515	0.493	0.468	-0.058
2015 CBA	Mathematics	0.426	0.416	0.411	0.403	-0.023
	Reading	0.587	0.548	0.554	0.522	-0.065
	Science Trend	0.493	0.465	0.476	0.452	-0.042
	Science New	0.459	0.428	0.445	0.415	-0.044
PISA-D PBA	Mathematics	0.296	0.290	0.278	0.257	-0.039
	Reading	0.410	0.391	0.364	0.329	-0.081
	Reading Components*	0.816	NA	0.807	NA	-0.010
	Science	0.385	0.382	0.366	0.354	-0.031

* For Reading Components, the difference is taken between position 1 and 3 because these reading components clusters were always placed in either Position 1 or Position 3.

The omission rates at different positions for all PISA-D countries were analysed to further examine the quality of data affected by position. The omission rates for the PISA-D Main Survey are shown in Table 9.7 for all domains and cluster positions. These rates do not include "not reached" items.

Table 9.7 **PISA-D PBA average omission rates across clusters and across countries**

	Position 1	Position 2	Position 3	Position 4	Position 4-Position 1*
Mathematics	0.112	0.108	0.127	0.150	0.038
Reading	0.095	0.105	0.124	0.161	0.065
Reading Components	0.025	NA	0.023	NA	-0.002
Science	0.070	0.070	0.093	0.113	0.043

The omission rates in positions 3 and 4 are higher than those in positions 1 and 2. This may be an indication that some students spent considerably more time on clusters 1 and 2, leaving less time for clusters 3 and 4.

Item correlations

The IRT models used for scaling assume conditional independence among items. If conditional independence does not occur for some of the items, the slope parameters for those items and the test reliability would be overestimated. To monitor the conditional dependencies among items, item-by-item correlations were examined. Relatively high item-by-item correlations would suggest that dependencies exist among those items, which may be caused by, for example, too much similarity among items (e.g., repetitive items), information in one item providing a clue to solving another, or information in the stimulus common to a set of items that is critical to correctly responding to more than one of the set items. When the local independence assumption is met, a similar level of correlation is expected among all the item pairs within a cluster, and no distinctive pattern can be discerned. In large-scale assessments, low to medium correlations are typically expected as items within a domain are expected to collectively contribute to the test information and to form a common scale. When interpreting the magnitude of correlations, within-cluster correlations can be compared against the across-cluster correlations. Given that data have dichotomous (0,1) or polytomous (0,1,2) scores, the 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.

Table 9.8 summarises the distribution of averages of item-by-item correlations across all countries: within the Main Survey cluster as well as across clusters for each domain (Mathematics, Reading, Reading Components, and Science). In each domain, there were no item pairs with problematic item correlations at each country level and across countries.

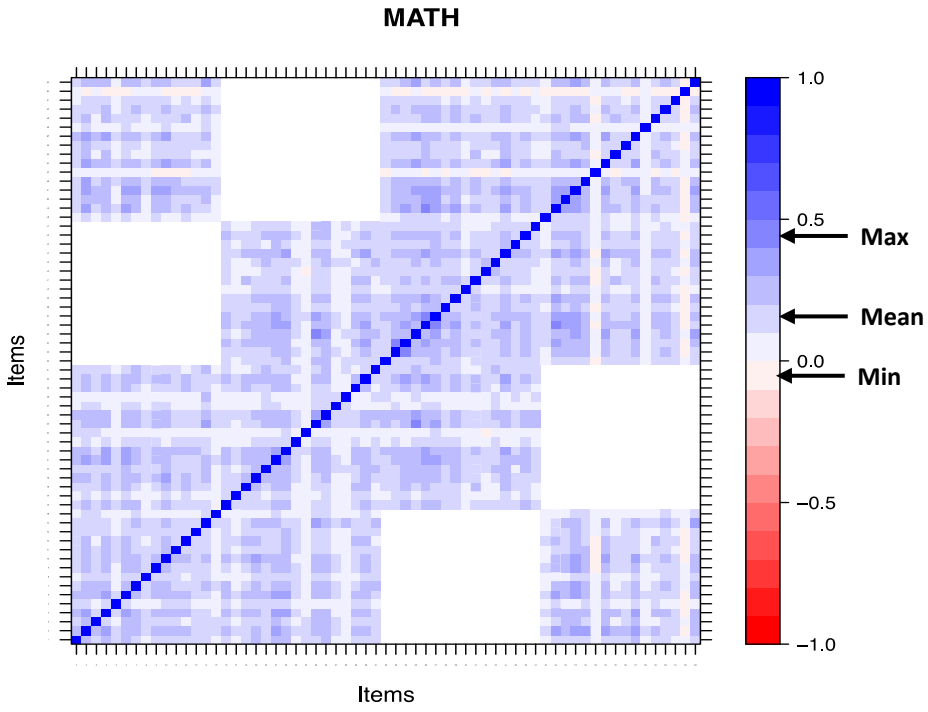
Table 9.8 **Distribution of averages of item-by-item correlations across all countries in each domain**

Domain	Cluster	Mean	Minimum	10th Percentile	90th Percentile	Maximum
Mathematics	M1	0.166	0.027	0.078	0.259	0.342
	M2	0.145	0.015	0.066	0.246	0.363
	M3	0.188	0.034	0.110	0.282	0.414
	M4	0.135	-0.035	-0.001	0.276	0.345
	Overall	0.156	-0.067	0.055	0.269	0.434
Reading	R1	0.153	0.001	0.062	0.236	0.530
	R2	0.167	0.057	0.081	0.261	0.413
	R3	0.167	0.013	0.069	0.260	0.347
	R4	0.169	0.024	0.071	0.278	0.374
	Overall	0.156	-0.009	0.068	0.244	0.530
Reading Components	RC1	0.195	-0.007	0.052	0.371	0.484
	RC2	0.176	-0.053	0.052	0.313	0.473
	RC3	0.132	-0.088	0.019	0.253	0.511
	RC4	0.176	0.000	0.054	0.300	0.378
	Overall	0.172	-0.088	0.047	0.316	0.511
Science	S1	0.122	0.004	0.057	0.191	0.307
	S2	0.093	0.012	0.047	0.153	0.284
	S3	0.098	-0.008	0.036	0.169	0.231
	S4	0.098	-0.017	0.035	0.163	0.386
	Overall	0.101	-0.034	0.037	0.176	0.386

Figures 9.2, 9.3, 9.4, and 9.5 visualise the item-by-item correlation for all items in each domain. These figures illustrate that similar correlations were observed within each cluster, and no cluster stood out in terms of dependencies among items. For Reading Components, relatively higher correlations are represented by darker shades. High correlations were not seen as problematic here, as they sometimes occur when very difficult or very easy items are located adjacently, and Reading Components items were designed to be very easy.

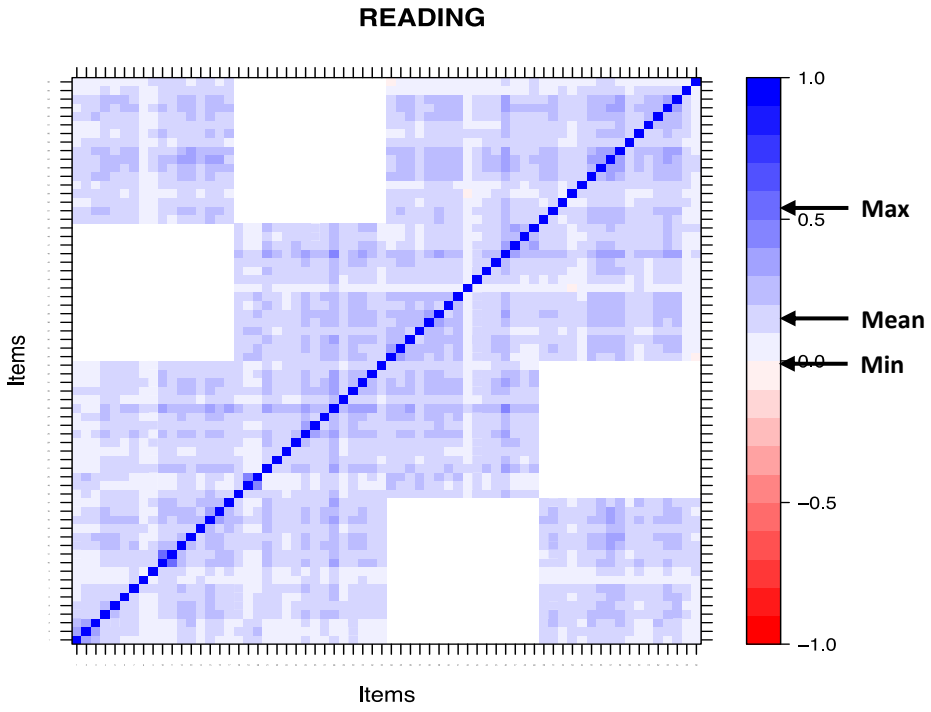
Figure 9.2

Item-by-item correlation for Math



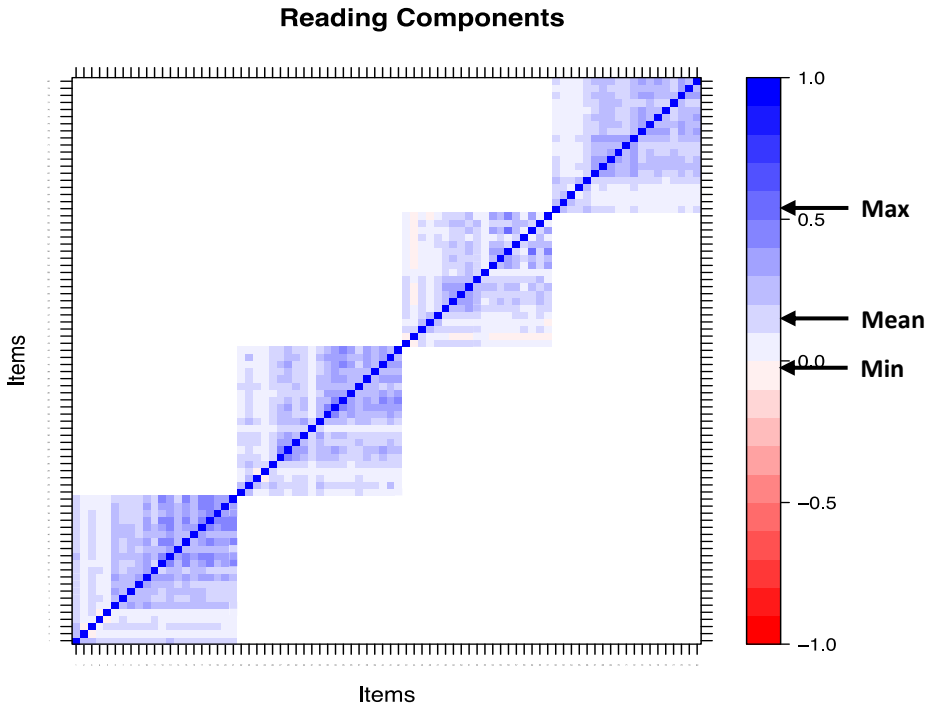
■ Figure 9.3 ■

Item-by-item correlation for Reading



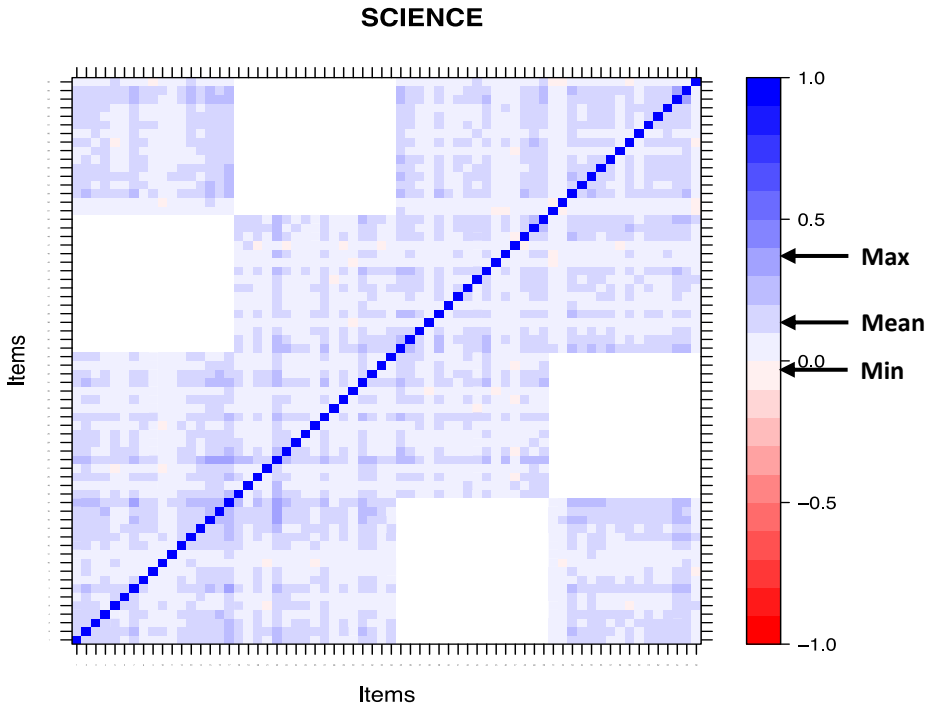
■ Figure 9.4 ■

Item-by-item correlation for Reading Components



■ Figure 9.5 ■

Item-by-item correlation for Science



SCALING METHODS IN PISA-D

This section describes the implementation of the different steps of IRT scaling and population modeling analyses of the PISA-D Main Survey data. First, the national and international item calibration is described. Then, the implementation of the population model and the computation of plausible values are described. In particular, the procedures utilised to link the scales to PISA and across PISA-D participating countries, are illustrated.

Scaling and analyses of the PISA-D data were carried out separately for each of the cognitive domains: Reading (together with Reading Components), Mathematics, and Science. By creating a separate scale for each domain, it remains possible to explore potential differences in subpopulation performance across these skills. The population model was carried out separately for each country.

The IRT models for scaling in PISA-D

The primary goal of the PISA-D scaling is to provide a reliable and valid link to the PISA 2015 scale so that PISA-D participating countries can be located on a comparable scale. For all three cognitive domains (Math, Reading and Reading Components, and Science) in the PISA-D Main Survey, comparability was established 1) to the PISA 2015 Main Survey, and 2) across the PISA-D participating countries.

The IRT scaling follows the procedures used in PISA 2015 and uses the same unidimensional IRT models: the two-parameter-logistic model (2PLM; Birnbaum, 1968) for dichotomously scored responses and the general partial credit model (GPCM; Muraki, 1992) for items with more than two ordered response categories.

The 2PLM is a generalisation of the Rasch model. Similar to the Rasch model, the 2PLM assumes that the probability of response *x* to item *i* by a respondent depends on the difference between the respondent's proficiency ϑ and the difficulty of the item difficulty, β_i. But in addition, the 2PLM allows the association between this difference and the response probability, for every item, to be able to depend on an additional item discrimination parameter (α_i), characterising the sensitivity of the item to proficiency. With the 2PLM, the response probability to an item is given as a function of this person parameter and the two item parameters, and it can be written as:

$$P(x_{ij} = 1|\theta, \beta_i, \alpha_i) = \frac{\exp(D\alpha_i(\theta - \beta_i))}{1 + \exp(\mathrm{D}\alpha_i(\theta - \beta_i))} \qquad (9.1)$$

where D is a constant of arbitrary sise, often either 1.0 or 1.7, depending on the parameterisation used in the software implementation. In the case of PISA-D, same as PISA 2015 Main Survey, a value of 1.7 is used. Note that, for $\alpha_i > 0.0$ this is a monotone increasing function with respect to θ; that is, the conditional probability of a correct response increases as the value of θ increases. One important special case is when $\alpha_i = 1.0/D$ for all items, in which case we can recognise the Rasch model as a special case of the 2PLM. This means that the 2PLM does not force a difference from the Rasch model; it only differs from the model if the optimal estimates for the slope parameter are different across items.

The GPCM (Muraki, 1992), like the 2PLM, is a mathematical model for responses to items with two or more ordered response categories. While the 2PLM is suitable for dichotomous responses only, the GPCM can be used with polytomous and dichotomous responses. The GPCM reduces to the 2PLM when applied to dichotomous responses. For an item *i* with m_i+1 ordered categories, the model equation of the GPCM can be written as:

$$P(x_i = k|\theta, \beta_i, \alpha_i, d_i) = \frac{\exp\{\sum_{r=1}^{k} \mathrm{D}\alpha_i\,(\theta - \beta_i + d_{ir})\}}{\sum_{u=0}^{m_i} \exp\{\sum_{r=1}^{u} \mathrm{D}\alpha_i\,(\theta - \beta_i + d_{ir})\}} \qquad (9.2)$$

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

As indicated earlier, 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 students, responses to any other items presented in a test, or the survey administration conditions. Another important assumption made that is consistent with the assessment framework is that the primary (often single) score for each domain measured can be accounted for by a dominant latent variable, ϑ, (unidimensionality). When these assumptions are satisfied, the unidimensional IRT models just described can be used. Then the joint probability of a particular response pattern $\boldsymbol{x} = (x_1, \ldots, x_n)$ across a set of n items can be expressed as follows.

$$P(\boldsymbol{x}|\theta, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^{n} P_i(\theta)^{x_i} \left(1 - P_i(\theta)\right)^{1-x_i} \qquad (9.3)$$

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 students provide their answers independently of one another and that the student's proficiencies are sampled from a distribution, $f(\theta)$. The likelihood function is therefore characterised as:

$$P(\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{j=1}^{J} \int \left(\prod_{i=1}^{n} P_i(\theta)^{x_{ij}} \left(1 - P_i(\theta)\right)^{1-x_{ij}} \right) f(\theta) d\theta \qquad (9.4)$$

Given a scored item response dataset and a choice of item response models (here a mixture of 2PLM and GPCM for dichotomous and polytomously scored items), the item parameters and the person latent traits can be estimated by maximising this function.

In order to ensure that the IRT models used provide adequate fit to the observed data, different types of model fit checks can be applied. One of these checks is the evaluation of differential item functioning (DIF) to determine whether, after taking differences in ability into account, items are harder or easier and/or more or less discriminating for a particular group when compared to the common or the fixed item parameters. More specifically, for each item, the empirical item characteristic curves (ICC) for each country were compared to the expected ICC given the common item parameter based on the total sample or the fixed parameters, as is the case for PISA-D. Noticeable differences between empirical and expected ICCs for a certain group or for all groups would be evidence of DIF. Following the same approach as used in PISA 2015, the mean deviation (MD) and the root mean square deviation (RMSD) indices were computed to quantify the magnitude and direction of DIF. While MD is sensitive to deviations in item difficulty, RMSD is sensitive to the deviations in both item difficulty and item.

Items in PISA-D that showed deviations from the common PISA item parameters were assumed to work differently in PISA-D and, therefore, to possibly harm the link to PISA. Items that showed deviations from the newly estimated common item parameters in PISA-D were assumed to work differently in certain countries. Poorly fitting items were identified using an RMSD > 0.15, and an |MD| > 0.15 criterion, where a value of 0 indicates no discrepancy (in other words, a perfect fit of the model). It was assumed for such items that the common item parameters were not appropriate; group-specific unique item parameters were estimated in a second step. Group-specific item parameters (i.e., national item parameters) for items exhibiting group-level DIF in the international calibration were estimated to reduce potential bias introduced by these deviations. The approach of using country-specific item parameters was favored over dropping the group-specific item responses for these items from the analysis. While the items with country DIF that were treated in this way no longer contribute to the international set of comparable responses, they continue to contribute to the reduction of measurement uncertainty for the specific country.

In the subsequent step, unique item parameters were estimated to account for country-specific deviations for a small subset of items. This involved a close monitoring of the IRT scaling for item-by-country interactions and allowing country-specific item parameters only in instances where substantial deviations were identified. This procedure takes measurement error into account, that is, it considers that some items work differently in certain countries. The common and unique item parameters were estimated using a mathematical algorithm that still allows us to estimate all item parameters in relation to one another, and thus, common and unique item parameters were on the same latent scale. Having a large number of common item parameters supports the comparability of the scales across the countries and assessments, while having only a few unique item parameters only reduces the measurement error further and does not affect the comparability of scales.

The software used for item calibration, *mdltm* (von Davier, 2005), implements an algorithm that monitored DIF measures, which automatically generated a suggested list of group-specific item treatments. This algorithm grouped similar deviations of subgroups so that unique parameters were assigned to either an individual country or multiple countries that showed the same level and direction of deviation.

Extensive descriptions of the methodologies and procedure are provided in the following references. The analyses of the PISA-D Main Survey and PISA 2015 Main Survey follow best practices outlined in, for example, 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). The methods used in PISA, as well as other assessments, 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 and Rabe-Hesketh, 2004; von Davier and Yamamoto, 2004, 2007; Adams, Wu, and Carstensen, 2007). The approach taken for the PISA 2015 analysis is a model that combines features of the Rasch model/PCM and the 2PLM/GPCM. This more general model was applied to the PISA 2015 Field Trial and Main Survey data. 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. The specific procedure used for PISA 2015 is described below in more detail. Based on the research studies just cited, the approach can be expected to help to retain linking items across modes or from prior assessments that would otherwise be excluded from the trend measure (the more linking items with good fit across groups, the more stable the link becomes).

Developing common scales between PISA-D and PISA

PISA-D was administered as a PBA linked to PISA, which means that a majority of items were selected from previous cycles of PISA but are complemented with existing materials from surveys including PISA for Schools, PIAAC, STEP, and LAMP. There were no new cognitive items developed for PISA-D. Within these clusters, at least half of the items were trend items from PISA. The rest were from different existing surveys different from PISA 2015.

Following the PISA 2015 Main Survey procedures, linking to PISA 2015 was established through multiple group IRT models (concurrent calibration) with fixed item parameter linking and the assumption of equal item parameters across groups. Linking items (trend PISA items for Mathematics, Reading, and Science as well as new Science items) were fixed to the common item parameters obtained from the PISA 2015 Main Survey to evaluate the functioning of items. While trend items were fixed to item parameters from the PISA 2015 PBA items, new Science items were fixed to item parameters from the PISA 2015 CBA items, since these items exist in computer-based format only in PISA 2015 (and were adjusted to a paper-based format for the PISA-D). For items coming from other (non-PISA) sources and items that were adapted for PISA-D (e.g., items that were changed to include partial credit score), item parameters were estimated. For these items, equality constraints were imposed so that common item parameters were estimated across the seven PISA-D countries. The scaling was carried out separately for each of the cognitive domains (Mathematics, Reading, and Science), and Reading Components items were scaled together with Reading items.

Table 9.9 gives an overview of the sources of items used in the PISA-D Main Survey cognitive assessment and how they were treated. PISA 2015 items (PBA items from PISA 2015 Mathematics, Reading, and Science, and a handful of CBA items from PISA 2015 Science) serve as linking items to construct a PISA-D scale that is comparable to the PISA 2015 scale. The PISA-D Main Survey cognitive assessment followed the same scoring guidelines and procedures as those applied in the PISA 2015 Main Survey for the paper-based assessment in order to maintain comparability between the two studies. Among the PISA trend items, partial credit scores were added (i.e., response categories from 2 to 3) for some of the items in Mathematics, Reading, and Science to obtain a more precise measurement of the lower end of the proficiency scale for PISA-D. These items for which partial credit scores were applied were not able to serve as linking items because the same item parameters as the PISA 2015 could not be used for scaling. Items from other sources (PISA for Schools, PIAAC, and LAMP) were not considered linking items. Taken together, in Table 9.6, linking items are shaded (i.e., source is “PISA 2015 Trend/New” and treatment is “fixed”), which constituted 49.2% in Mathematics, 77.3% in Reading, and 65.2% in Science.

The items for which partial credit scores were added are listed in Table 9.10. In the domain of Science, two more PISA 2015 New items were not fixed but estimated because those two New items were excluded in the PISA 2015 Main Survey data analysis; thus, no item parameters were available to be fixed in the scaling.

Table 9.9 **Items in PISA-D by source and treatment**

Mathematics	Link/Anchor Items	Estimated	Total
PISA 2015 PBA	31	8	39
PISA for Schools		11	11
PIAAC		13	13
Total	31	32	63
Reading	**Link/Anchor Items**	**Estimated**	**Total**
PISA 2015 PBA	51	2	53
PISA for Schools		4	4
LAMP		5	5
PIAAC		4	4
Total	51	15	66
Science	**Link/Anchor Items**	**Estimated**	**Total**
PISA 2015 PBA	37	8	45
PISA 2015 CBA	6	4	10
PISA for Schools		11	11
Total	43	23	66

Table 9.10 **Items for which partial credit scores were added for the PISA-D**

Domain	Item	Item Format
Math (8 items)	PM192Q01S	Complex Multiple Choice
	PM464Q01S	Open Response - Human Coded
	PM948Q03A	Open Response - Human Coded
	PM496Q01S	Complex Multiple Choice
	PM273Q01S	Complex Multiple Choice
	PM919Q01A	Open Response - Human Coded
	PM949Q01S	Complex Multiple Choice
	PM411Q01A	Open Response - Human Coded
Reading (2 items)	PR404Q07AS	Complex Multiple Choice
	PR432Q06AS	Complex Multiple Choice
Science (10 items)	PS638Q02AS*	Complex Multiple Choice
	PS638Q04S*	Complex Multiple Choice
	PS415Q08S	Complex Multiple Choice
	PS498Q02S	Complex Multiple Choice
	PS413Q04S	Complex Multiple Choice
	PS466Q01S	Complex Multiple Choice
	PS478Q02S	Complex Multiple Choice
	PS527Q01S	Complex Multiple Choice
	PS527Q04S	Complex Multiple Choice
	PS408Q04S	Complex Multiple Choice

* Two science items were newly developed for the PISA 2015 Main Survey; thus, these CBA (computer-based assessments) items were adapted to PBA items for PISA-D and administered as PBA items.

National and international item calibration

Because the samples for the PISA-D Main Survey came from populations with somewhat different characteristics, the calibration procedure needed to take into account the possibility of interactions between the samples and the items that were used to produce estimates of the item parameters. For this reason, a multiple-group IRT model, treating each country as a distinctive group, was estimated using a mixture of normal population distributions (one for each sample) where item parameters were generally constrained to be equal across groups with a unique mean and variance for each country. The moments of these distributions (i.e., mean and variance) were updated for every step in the iterations of the item parameter estimation.

The item calibration was completed in two consecutive steps. First, linking to PISA 2015 was established through multiple-group IRT models with fixed item parameter linking (concurrent calibration) and assumed equal item parameters across groups for estimating new item parameters. More specifically, linking items (trend PISA items for Mathematics, Reading, and Science, as well as New Science items) were fixed to the common item parameters obtained from the PISA 2015 Main Survey to evaluate item functioning.

In the subsequent step, unique item parameters were estimated to account for country-specific deviations for a small subset of items (see the next subsection for details on country-specific item parameter determination). This involved a close monitoring of the IRT scaling for item-by-country interactions and allowing country-specific item parameters only in instances where substantial deviations were identified. This procedure takes measurement error into account, that is, it considers that some items work differently in certain countries. For items coming from non-PISA sources and for items that were adapted for PISA-D (e.g., items that were changed to include partial credit score), equality constraints were imposed so that common item parameters were estimated across the seven PISA-D countries, and the scaling was carried out separately for each of the cognitive domains. The common and unique item parameters were estimated using a mathematical algorithm that still allowed us to estimate all item parameters in relation to one another, and thus, common and unique item parameters were mapped onto the same latent scale. Having a large number of common item parameters supports the comparability of the scales across the countries and assessments, while having only a few unique item parameters reduces the measurement error further and does not affect the comparability of scales.

For the domain of Reading, additional steps have been followed to establish a common scale together with Reading Components items. Reading Components items were introduced in PISA-D to better describe lower end proficiency. First, only Reading items (66 items) were scaled fixing linking item parameters to the common item parameters obtained in PISA 2015 Main Survey. Second, Reading items were finalised by evaluating the item fit statistics and by allowing a small number of country-specific item parameters if needed. This is to ensure that the Reading scale is not to be affected by Reading Components items. Lastly, Reading Components items were added to the finalised Reading scale and similar steps were followed to allow unique item parameters for Reading Components item parameters. Those steps were as follows: Start with common item parameters for all Reading Components items, evaluate the item fit statistics for each item-by-country combinations, and allow small number of country-specific item parameters if needed.

In PISA-D, as with PISA 2015, omitted responses prior to a valid response were treated as incorrect responses because a random response to an open-ended item would almost certainly result in a wrong answer; in contrast, omitted responses at the end of each of the two one-hour test sessions were treated as not reached/not administered. In the latter case, impact on the IRT scaling was avoided by excluding these responses when the likelihood function was calculated. However, the number of not-reached items was introduced as a covariate in the latent regression model, so it is part of the proficiency estimation in the generation of plausible values (see the section titled "Latent regression model for population modeling").

Handling of item-by-country interactions

Given that international assessments are translated into multiple target languages, item-by-country interactions are a potential threat to validity (e.g., some terms may be harder to translate into a specific target language, and, in the process, the way or the content the source item measures may be altered). As such, some items in some countries may function somewhat differently from how the item generally functions in the majority of countries or groups, or how the item generally functions in the majority of participating countries. Therefore, the consistency of item parameter estimates across countries was of particular interest to achieve equivalent and comparable measurement across countries as well as between the PISA and PISA-D scales.

If a test measures the same latent trait in a given domain in all groups, the items should have the same relative difficulty or, more precisely, would fall within the interval defined by the standard error on the item parameter estimate (i.e., the confidence interval). In cases where common item parameters are not appropriate for certain items in certain groups (item-by-country interactions) as determined by group-specific item-fit statistics (MD and RMSD), unique item parameters were estimated in a stepwise procedure.

Given specifications (minimum sample size and maximum and minimum threshold values), items requiring unique parameters based on DIF detection were automatically identified by the *mdltm* software. If an item was identified for DIF and more than one group deviated from the international/common parameters in the same way (showing DIF in the same direction), the algorithm assigned item parameters common to those groups, but different from the international parameters. For example, if two groups (e.g., two countries in PISA-D) showed poor item fit for the same item in the international/common calibration, and in the same direction, both groups received the same unique item parameter estimated for these two groups but different from the rest of the groups (note that the term "unique item parameters" in this report is used for both cases: one group that receives a unique group-specific item parameter, and more than one group that receives the same unique item parameter that is different from the international/common item parameter). If an item showed poor fit to a different direction in different groups (e.g., negative MD vs. positive MD), unique group-specific item parameters were used for further analysis. Thus, PISA-D allowed for different sets of item parameters to improve model fit and optimise the comparability of groups and countries.

To identify misfitting items, fit statistics were estimated using the MD and the RMSD (see section titled "The IRT models for scaling in PISA-D" for more information on these statistics). Poorly fitting

items were identified using an RMSD > 0.15 criterion and an |MD| > 0.15 criterion (a value of 0 indicates no discrepancy; in other words, a perfect fit of the model). The identification of poorly fitting items and the replacement of international item parameters with group-specific (unique) parameters was carried out using an automatic algorithm in *mdltm*. Thus, the international and national calibrations were conducted simultaneously for all groups, so all of the estimated item parameters (international and unique) are located on a common scale.

Typically, only a small number of unique item parameters are assigned. The vast majority of items are expected to fit well for all, or nearly all, countries using international/common item parameters. Chapter 12 provides an overview of the percentage of group-specific item parameters per country.

LATENT REGRESSION MODEL FOR POPULATION MODELING

This section reviews the population (or conditioning) model—a combination of an IRT model and a latent regression model—employed in the analyses of the PISA-D data and explains the multiple imputation or "plausible values" methodology that aims to increase the accuracy of the estimates of the multivariate proficiency distributions for various subpopulations and the population as a whole.

Individual cognitive skills tests are concerned with accurately assessing the performance of individual students for the purposes of diagnosis, selection, or placement. The accuracy of these measurements can be improved (i.e., reducing the amount of measurement error) by increasing the number of items administered to the individual that measure the same skill. Thus, individual achievement tests containing more than 70 items are common. Because the uncertainty associated with each estimated proficiency θ is negligible, the distribution of proficiency or the joint distribution of proficiency with other variables can be approximated using individual proficiency estimates. But when analysing the distribution of proficiencies for populations or subpopulations, more efficient estimates can be obtained from a matrix-sampling design.

In international large-scale assessments (ILSAs) such as PISA, test forms are kept relatively short to minimise individuals' response burden. This is important since ILSAs are low-stakes assessments that do not provide feedback and do not entail consequences for the individual test taker. At the same time, ILSAs aim to achieve broad coverage of the tested constructs. In this context, the full set of items is organised into different, but linked, test forms; each individual receives only one form. Thus, the survey solicits relatively few responses from each student on any one domain while maintaining a wide range of content representation when responses are aggregated. The advantage of estimating population characteristics more efficiently is offset by the inability to reliably measure and make precise statements about individuals' performance on a single domain. As a consequence, point estimates of proficiency that are (in some sense) optimal for each student could lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, and Beaton, 1987).

In the case of ILSAs, improved proficiency distributions are derived that are based on both the relatively small number of responses to items in the booklet, and responses to background questions administered in the student questionnaire. In addition, the covariance between skill

domains (e.g., the PISA core domains, Mathematics, Reading, and Science) is utilised to further improve the estimates of skill distributions. This approach allows for estimation of proficiency distributions given responses received on all domains in the test booklet and the student questionnaire. The "plausible value" methodology uses these proficiency distributions and accounts for error (or uncertainty) at the individual level by using multiple imputed proficiency values (plausible values) rather than assuming that this type of uncertainty is zero. Retaining this component of uncertainty requires that additional analysis procedures be used to estimate student proficiencies. The population model used for the PISA-D is essentially same as the PISA 2015 Main Survey: incorporated test responses (responses to the cognitive items) as well as variables measured by the student context questionnaire (e.g., academic and nonacademic activities, and attitudes), which serve as covariates, in the computation of plausible values (von Davier, Sinharay, Oranje, and Beaton, 2006). Ten plausible values are randomly selected for each student. The combined model requires the estimation of the IRT measurement model, which provides information about test performance, and the latent regression, which provides information about the extent to which student background information, can predict proficiency. The estimation of this combined model is carried out as follows:

1. *Item calibration based on IRT* (scaling): The responses consist of dichotomously and polytomously scored values. These responses are used to calibrate the test and provide item parameter estimates for the (cognitive) test items. The 2PLM is fitted for dichotomous item responses and the GPCM is fitted for polytomous item responses. Note that for a subset of trend items, the Rasch model and the PCM continue to be fitted for dichotomous and polytomous responses, respectively, to maintain consistency with prior PISA cycles.
2. *Population modeling using latent regressions:* The population model assumes that item parameters are fixed at the values obtained in the calibration stage. Taking the item parameters estimates from item calibration, a latent regression model is fitted to the data to obtain regression weights (Γ) and a residual variance-covariance matrix for the latent regression ($\sum$).
3. *Plausible value generation*: Ten plausible values (Mislevy and Sheehan, 1987; von Davier, Gonzalez and Mislevy, 2009) are drawn for all students using the item parameter estimates from the item calibration stage and the estimates of Γ and $\sum$ from the latent regression model.

In the latent regression model, the distribution of the proficiency variable, ϑ, is assumed to depend on the cognitive item responses, *X,* as well as background variables, *Y*, derived from responses obtained from the context questionnaire (e.g., gender, country of birth, reading practices, etc.). The item parameters from the calibration stage and the estimates from the regression analysis are both needed to generate plausible values.

A considerable number of background variables (predictors) are usually collected in ILSAs. Principal components accounting for a large proportion of the variation in the context questionnaire variables were used in the latent regression instead of the observed context questionnaire variables. The use of principal components serves to retain information for students with missing responses to one or more background variables. For PISA-D, components for each country that accounted for 80% of the variance were used in order to avoid numerical instability due to potential overparameterisation of the model. If the number of principal components that explain 80% of the

variance exceeded a certain threshold per country (i.e., 5% of raw sample size), it was chosen to use the number of principal components based on the 5% of the sample size. This was done to explain as much variance as possible while at the same time avoiding overparameterisation of the model. For the regression of the background variables on the proficiency variable it was assumed that:

$$\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{y}\Gamma, \sum) \tag{9.5}$$

The latent regression parameters Γ and $\sum$ were estimated conditional on the previously determined item parameter estimates (from the item calibration stage). Γ is the matrix of regression coefficients and Σ is a common residual variance-covariance matrix.

The latent regression model of Θ on Y with $\Gamma = (\gamma_{sl}, s = 1,\ldots,S; l = 0,\ldots,L)$, $Y = (1, y_1, \ldots, y_L)^t$, and $\Theta = (\theta_1, \ldots, \theta_S)^t$ can be described as follows:

$$\theta_s = \gamma_{s0} + \gamma_{s1}y_1 + \ldots + \gamma_{sL}y_L + \varepsilon_s \tag{9.6}$$

where ε_s is an error term for the assessment skill s.

The residual variance-covariance matrix can then be estimated using the equation:

$$\Sigma = \Theta\Theta^t - \Gamma(YY^t)\Gamma^t \tag{9.7}$$

Plausible values for each student *j* are drawn from the conditional distribution:

$$P\left(\theta_j \middle| \boldsymbol{x}_j, \boldsymbol{y}_j, \Gamma, \Sigma\right) \tag{9.8}$$

Using standard rules of probability, the conditional probability of proficiency can be represented:

$$P\left(\theta_j \middle| \boldsymbol{x}_j, \boldsymbol{y}_j, \Gamma, \Sigma\right) \propto P\left(\boldsymbol{x}_j \middle| \theta_j, \boldsymbol{y}_j, \Gamma, \Sigma\right) P(\theta_j | \boldsymbol{y}_j, \Gamma, \Sigma) \; = P\left(\boldsymbol{x}_j \middle| \theta_j\right) P(\theta_j | \boldsymbol{y}_j, \Gamma, \Sigma) \tag{9.9}$$

where ϑ_j is a vector of scale values (these values correspond to performance on each of the skills), $P(\boldsymbol{x}_j|\vartheta_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\vartheta_j|\boldsymbol{y}_j, \Gamma, \Sigma)$ is the multivariate joint density of proficiencies of the scales, conditional on the principal components y_j derived from background responses, and parameters Γ and Σ. The item parameters are fixed and regarded as population values in the latent regression modeling stage.

The basic method for estimating Γ and Σ using the expectation-maximisation (EM) algorithm is described in Mislevy (1985) for the single scale case. The EM algorithm requires the computation of the mean and variance of the posterior distribution in the equation above.

After the estimation of Γ and Σ is complete, plausible values are drawn from the joint distribution of the values of Γ for all sampled students in a three-step process. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma|\boldsymbol{x}_j,\boldsymbol{y}_j)$ that fixes Σ at the value $\hat{\Sigma}$ (Thomas, 1993). Second,

conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean m_j^p, and variance Σ_j^p of the posterior distribution are computed using the same methods applied in the EM algorithm. In the third step, the ϑ are drawn independently from a multivariate normal distribution with mean vector m_j^p and posterior co-variance matrix Σ_j^p. These three steps were repeated 10 times, producing 10 imputations of ϑ for each sampled student.

The software DGROUP (Rogers, Tang, Lin, and Kandathil, 2006) was used to estimate the latent regression model and generate plausible values (von Davier, Sinharay, Oranje, and Beaton, 2006; von Davier and Sinharay, 2014). A multidimensional variant of the latent regression model based on Laplace approximation (Thomas, 1993) was applied, as PISA reports proficiencies on more than two skill dimensions.

Population modeling in PISA-D

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

As in the PISA 2015 Main Survey, a minimum of six completed items per domain was necessary to assure sufficient information about the proficiency of students. In general, there were very few students[2] (0.04%) with responses to fewer than six cognitive items in at least one of the main cognitive domains. Thus, this two-step procedure was taken: In the first round, respondents who responded to at least six items within at least one domain were used to fit the multidimensional latent regression models when Γ and Σ were estimated, and in the second round, all respondents, including those who responded less than six items, received plausible values fixing the regression parameters to the ones obtained from the first run. This procedure ensured that the cases with fewer item responses did not contribute to the estimation of the proficiency distribution, but did receive the plausible values in the domain that they responded to.

Generating plausible values

In PISA-D, the computation of group-level reporting statistics involving scores in the main cognitive domains was based on 10 independently drawn plausible values for each of the cognitive domains for each student. Each set of plausible values was equally well designed to estimate population parameters; however, multiple plausible values were required to represent the uncertainty in the domain measures appropriately (von Davier, Gonzalez, and Mislevy, 2009). The statistics based on scores are always computed at population or subpopulation levels. They should never be used to draw inferences at the individual level. Detailed information on the computation and the use of plausible values in analyses is given in Rutkowski, Gonzalez, Joncas, von Davier (2010).

[2] Note that a student was only considered a "respondent" and given an analysis weight if he or she responded to at least one cognitive item and possessed data for the context questionnaire items, or if he or she responded to at least half of the cognitive items in cases of providing no context questionnaire information.

For population modeling and for generating plausible values for three scales of PISA-D, the computer program DGROUP (Rogers et al., 2006) was used. A normal multivariate distribution was assumed for $P(\vartheta_j | x_j, y_j, \Gamma, \Sigma)$, with a common variance, Σ, and with a mean given by a linear model with slope parameters, Γ, based on the principal components of several hundred selected main effects from the vector of context questionnaire variables.

The background variables included nearly all student questionnaire data, school ID, gender, and the number of not-reached items, among others. A description of the different sections of the background data can be found in Chapter 3. All variables in the context questionnaire were contrast coded before they were processed further in the principal components analysis. Contrast coding allows for the inclusion of codes for refused responses, avoiding the necessity of linear coding. However, the increased number of variables obtained through contrast coding is substantial. To capture most of the common variance in the contrast-coded background questions with a reduced set of variables, a principal component analysis was conducted. Because each population can have unique associations among the background variables, a single set of principal components was not sufficient for all countries included in PISA-D. As such, the extraction of principal components was carried out separately by country to take into account the differences in associations between the background variables and the cognitive skills. The plausible value variables for the cognitive domains follow the naming convention PV1xxxx through PV10xxxx, where "xxxx" takes on the following form:

- READ for Reading Literacy
- MATH for Mathematics Literacy
- SCIE for Science Literacy

Students received plausible values for each cognitive domain administered in their country according to the test design that was applied in a particular country. This means that students also received plausible values for cognitive domains that were not administered to them.

References

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

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

Birnbaum, A. (1968), "Some latent trait models and their use in inferring a student's ability, In F. M. Lord and M. R. Novick (eds.)", *Statistical Theories of Mental Test Scores,* Addison-Wesley, Reading, MA.

Glas, C. A. W. and **K. Jehangir** (2014), "Modeling country specific differential item functioning", In L. Rutkowski, M. von Davier, and D. Rutkowski (eds.), *Handbook of International Large-Scale Assessment,* CRC Press, Boca Raton, FL.

Glas, C. A. W. and **N. D. Verhelst** (1995), "Testing the Rasch model", In G. H. Fischer and I. W. Molenaar (eds.), *Rasch models: Foundations, Recent Developments, and Applications* (pp. 69-95)", Springer, New York, NY.

Mazzeo, J. and **M. von Davier** (2014), "Linking scales in international large-scale assessments", In L. Rutkowski, M. von Davier, and D. Rutkowski (eds.)", *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, CRC Press, Boca Raton, FL.

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

Mislevy, R. J. and **K. M. Sheehan** (1987), "Marginal estimation procedures", In A. E. Beaton (ed.), *Implementing the New Design: The NAEP 1983-84 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, NJ.

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

Oliveri, M. E. and **M. von Davier** (2011), "Investigation of model fit and score scale comparability in international assessments", *Psychological Test and Assessment Modeling,* 53(3), pp. 315-333, retrieved from *http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf*.

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

Rogers, A., C. Tang, M.-J. Lin and **M. Kandathil.** (2006), DGROUP (computer software), Educational Testing Service, Princeton, NJ.

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

Skrondal, A. and S. Rabe-Hesketh (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Chapman and Hall/CRC, Boca Raton, FL.

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

von Davier, M. (2005), *A General Diagnostic Model Applied to Language Testing Data* (Research Report No. RR-05-16), Educational Testing Service, Princeton, NJ.

von Davier, M., E. Gonzalez and **R. Mislevy.** (2009), "What are plausible values and why are they useful"? In *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, Vol. 2*.

Retrieved from IERI website, *http://www.ierinstitute.org/IERI_ Monograph_Volume_02_Chapter_01.pdf*.

von Davier, M. and **S. Sinharay** (2014), "Analytics in international large-scale assessments: Item response theory and population models", In L. Rutkowski, M. von Davier, and D. Rutkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis,* CRC Press, Boca Raton, FL.

von Davier, M. S., Sinharay, A. Oranje and **A. Beaton** (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*, Elsevier, Amsterdam, Netherlands.

von Davier, M. and **K. Yamamoto** (2004), "Partially observed mixtures of IRT models: An extension of the generalized partial credit model", *Applied Psychological Measuremen*t, 28(6), pp. 389-406.

von Davier, M. and **K. Yamamoto** (2007), "Mixture distribution Rasch models and Hybrid Rasch models", Chapter 6, in M. von Davier and C.H. Carstensen, *Multivariate and Mixture Distribution Rasch Models*, New York, NY, Springer.

Weeks, J., K. Yamamoto and **M. von Davier** (2014), "Design considerations for the Program for International Student Assessment", In L. Rutkowski, M. von Davier, and D. Rutkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, CRC Press, Boca Raton, FL.

Wingersky, M., B. Kaplan and **A. E. Beaton** (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, Princeton, NJ.

Yamamoto, K. (1997), "A chapter: Scaling and scale linking", *International Adult Literacy Survey Technical Report*, Statistics Canada, Ottawa, Canada.

Yamamoto, K. and **J. Mazzeo** (1992), "Item response theory scale linking in NAEP", *Journal of Educational Statistics,* Vol. 17/2, pp. 155-174.