

# Chapter 12: SCALING OUTCOMES

---

This chapter presents the outcomes of applying the item response theory (IRT) scaling and population models to the PISA-D Main Survey assessment data. Outcomes include: (i) the estimates of the items parameters; (ii) the generated plausible values and the reliability of the assessment, and (iii) the international characteristics of the item pool, the range of proficiencies targeted, and the domain inter-correlations.

## RESULTS OF THE IRT SCALING AND POPULATION MODELING

The assessment and linking design for PISA-D aimed to provide appropriate measurement for participating countries and establish a valid link to PISA 2015, placing participating countries on a comparable scale. For all three cognitive PISA-D domains (Mathematics, Reading, and Science), comparability was established (1) to the PISA 2015 Main Survey and (2) across the seven participating countries. As elaborated in Chapter 9 in the “Scaling Methods in PISA-D” section, when feasible, item parameters were either fixed to those obtained from the PISA 2015 Main Survey or were otherwise estimated with PISA-D data. Generally, items fit well among PISA-D participating countries and across PISA-D and PISA data, providing reliable measurement, comparability across countries, and strong linkage to the PISA 2015 Main Survey scales based on the common items between the two. However, there were a few items that functioned differently and some cases where the international item parameters did not fit well for a particular country or subset of countries. In these instances of item-by-country interaction, constraints on equal item parameters were released to allow the estimation of unique item parameters.

### Unique item parameter estimation and national item deletion

The IRT calibration for the PISA-D Main Survey data was carried out separately for each of the cognitive domains, Mathematics, Reading,<sup>1</sup> and Science. Because of either an item location issue (i.e., one Reading item located in Math cluster)<sup>2</sup> or having negative slope parameters consistently estimated in all countries, one Mathematics and two Reading items were excluded from the IRT analyses in all countries; Table 12.1 notes these items.

---

<sup>1</sup> Reading Components items were scaled together with Reading items to establish a unified scale for Reading.

<sup>2</sup> In one Math block, most items associated with a particular passage were, as one would expect, related to Math. However, one was actually a Reading item. This presented no scaling issue for students who took other Reading items as there was enough Reading data overall. But due to the spiral design of the assessment, many students did not take any other Reading items. For there to be data for only one Reading item for this set of students created instability in the IRT scaling. Therefore, the item was removed from the scale.

Table 12.1 **Items that were excluded from the IRT scaling**

Domain	Item	Reason
Math (1 item)	PM982Q04S	Negative slope estimate
Reading (1 item)	PR302PC02S	Located in Math cluster
Reading Components (1 item)	PR913P011S	Negative slope estimate

For the items sourced from non-PISA surveys in each cognitive domain, the unidimensional multigroup IRT analyses (detailed in Chapter 9) were conducted to estimate international/common parameters and unique national parameters while confirming the fixed PISA 2015 parameters for the linking items. Items that had good fit across all countries and did not require any unique national item parameters are called *scalar invariant items*. Scalar invariant items have the same slope and threshold parameters across countries. Scalar invariant items can be divided into two groups, depending on their source: PISA items and non-PISA items. The PISA scalar invariant items provide stable and strong linkage between PISA and PISA-D, and the full set of scalar invariant items (both PISA and non-PISA items) contribute to stable and strong linkage among PISA-D participating countries. The use of unique parameters reduces country-level measurement error without introducing a bias.

Table 12.2 shows the percentage of common and unique item parameters by domain computed by dividing the number of possible combinations of item-by-country pairs. Note that the percentages of scalar invariant international/common item parameters (i.e., sum of the first two rows) are all above 90% in all cognitive domains: 93.78% in Mathematics, 93.94% in Reading, and 95.45% in Science. More specifically, the first row directly contributes to PISA scale linkage as well as to comparability among PISA-D countries, and the second row contributes to the comparability among participating PISA-D countries. Introducing partial credit scores for some PISA linking items made it impossible to use the same item parameters in PISA-D. Thus, not all PISA items were used for linking. Because adapting items to have partial credit happened more often in Mathematics, the percentage in the first row for Mathematics is lower than 50%.

Further, only a small number of items received unique item parameters (either the same parameters across a subset of groups in the third row or a specific group in the fourth row) except for Reading Components. Reading Components items are considered more sensitive to a particular characteristic of languages such as orthography; thus, more unique item parameters were estimated across a subset of countries (0.90%) or to one specific country (5.06%). Group-specific unique item parameters in the fourth row showed similar proportions in Mathematics, Reading, and Science: about 2.6% to 2.8%.

Table 12.2 **Percentage of common and unique item parameters in each domain for PISA-D**

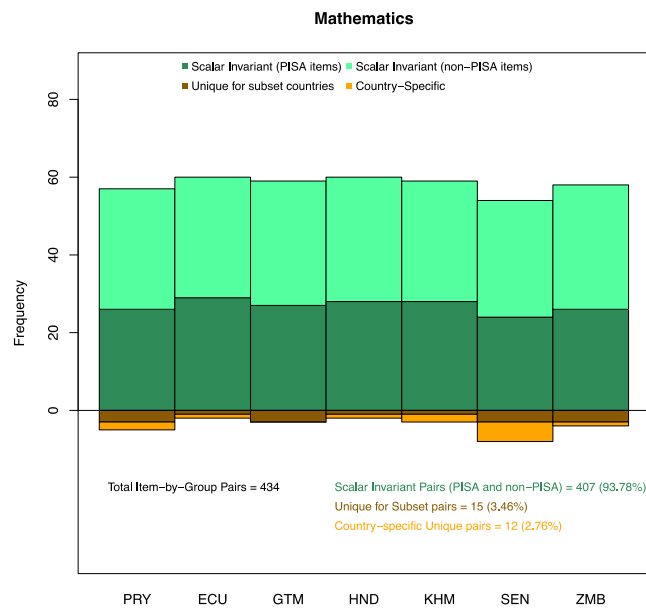
	<b>Mathematics</b>	<b>Reading</b>	<b>Reading Components</b>	<b>Science</b>
% of scalar invariant PISA common/international item parameters	43.32%	71.65%	NA	59.52%
% of scalar invariant non-PISA <sup>3</sup> common/international item parameters	50.46%	22.29%	94.04%	35.93%
% of unique item parameters (across a subset of groups)	3.46%	3.46%	0.90%	1.73%
% of unique item parameters (group-specific)	2.76%	2.60%	5.06%	2.81%
Number of items included in the PISA-D Main Survey	62	66	79	66

Proportions of international/common (invariant) item parameters and group-specific item parameters in each domain are visualised in Figures 12.1 to 12.4.

<sup>3</sup> These “non-PISA” items include PISA 2015 items for which partial credit scores were added; thus, fixed item parameter linking was not applicable, and item parameters were newly estimated.

■ Figure 12.1 ■

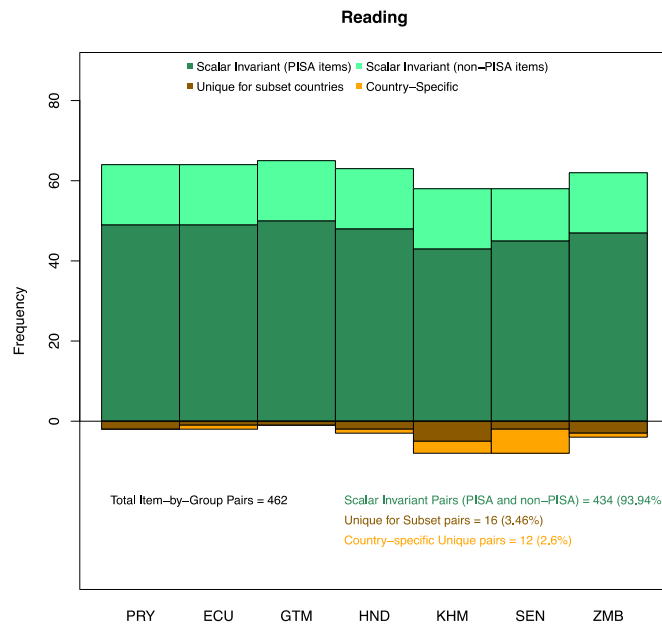
Frequencies of international and unique item parameters in Mathematics



Abbreviations: ECU = Ecuador, GTM = Guatemala, HND = Honduras, KHM = Cambodia, PRY = Paraguay, SEN = Senegal, ZMB = Zambia.

■ Figure 12.2 ■

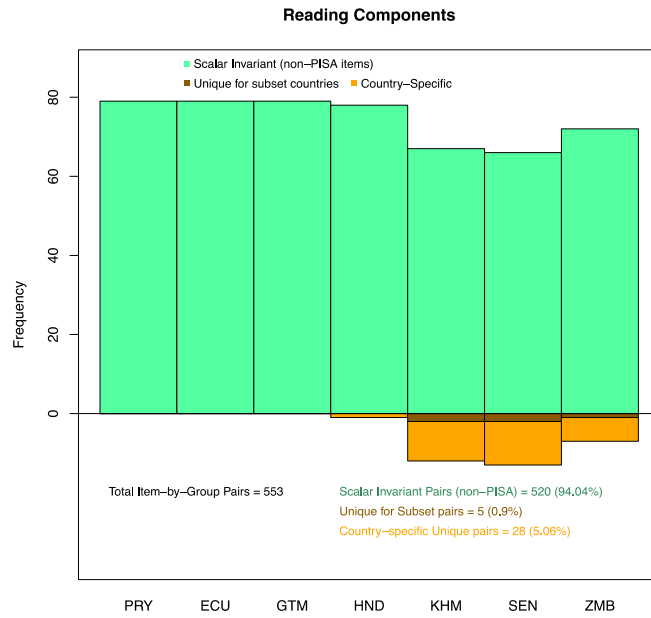
Frequencies of international and unique item parameters in Reading



Abbreviations: ECU = Ecuador, GTM = Guatemala, HND = Honduras, KHM = Cambodia, PRY = Paraguay, SEN = Senegal, ZMB = Zambia.

■ Figure 12.3 ■

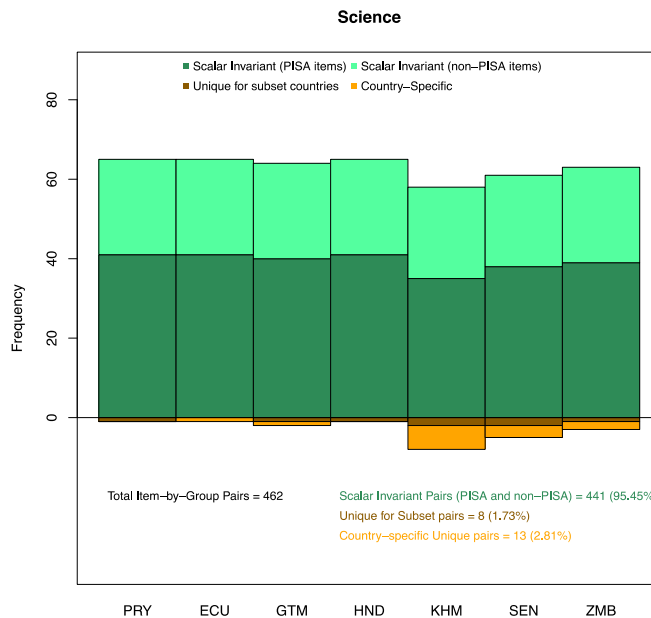
**Frequencies of international and unique item parameters in Reading Components**



Abbreviations: ECU = Ecuador, GTM = Guatemala, HND = Honduras, KHM = Cambodia, PRY = Paraguay, SEN = Senegal, ZMB = Zambia.

■ Figure 12.4 ■

**Frequencies of international and unique item parameters in Science**



Abbreviations: ECU = Ecuador, GTM = Guatemala, HND = Honduras, KHM = Cambodia, PRY = Paraguay, SEN = Senegal, ZMB = Zambia.

In each figure, green bars indicate the frequency of scalar invariant item parameters per country; linking items from PISA are dark green and non-PISA items are light green. The PISA items for which partial credit scores were added to better describe lower proficiency are considered to be non-PISA items in this count. Scalar invariant item parameters are visualised above the horizontal reference line, indicating that they contribute to the comparability among PISA-D countries and link to the PISA scales. Orange colors below the horizontal reference line indicate the frequency of unique item parameters; partial invariant item parameters that are common to a subset of countries but different from most countries are dark orange, and group-specific item parameters that contribute to the reduction of measurement error are light orange.

Results show that the overall item fit across countries is very good, resulting in a small number of unique item parameters and scores being highly comparable across PISA-D countries. There was no consistent pattern of deviations for any one particular country. The good fit of the PISA items supports strong linking to the PISA scales in each domain. An overview of the deviations per item across all countries for items in each domain is given in the Appendix of this chapter.

### **Generating student scale scores and reliability of the PISA-D scales**

Given the rotated and incomplete matrix assessment design, it is not possible to calculate marginal reliabilities for each cognitive domain in the classical sense. In order to estimate test reliability, the explained variance (i.e., variance explained by the model) for each cognitive domain was computed based on the weighted posterior variance. The explained variance was computed using all 10 plausible values with the equation:  $1 - (\text{expected error variance} / \text{total variance})$ . The weighted posterior variance is an expression of the posterior measurement error and is obtained through population modeling. The expected error variance is the weighted average of the posterior variance. This term was estimated using the weighted average of the variance of the plausible values (the posteriori variance is the variance across the 10 plausible values). The total variance was estimated using a resampling approach (Efron, 1982) and was estimated for each country depending on the country-specific proficiency distributions for each cognitive domain.

Applying the population model described in Chapter 9 and anchoring all of the item parameters at the values obtained from the final IRT scaling, plausible values were generated for all sampled students. Table 12.3 gives the median of national reliabilities for the generated scale scores based on all 10 plausible values. National reliabilities of the main cognitive domains are presented in Table 12.4.

Table 12.3 **Reliabilities of the PISA cognitive domains and Science subscales overall countries**

Domains	Median	S.D.	Max	Min
Mathematics	0.86	0.06	0.88	0.71
Reading <sup>4</sup>	0.88	0.06	0.89	0.72
Science	0.86	0.05	0.87	0.72

Table 12.4: **National reliabilities for main cognitive domains**

Country	Mathematics	Reading	Science
Ecuador	0.85	0.88	0.86
Guatemala	0.88	0.89	0.87
Honduras	0.88	0.87	0.86
Cambodia	0.82	0.84	0.77
Paraguay	0.86	0.88	0.87
Senegal	0.71	0.72	0.72
Zambia	0.86	0.89	0.83

Table 12.4 shows that the explained variance by the combined IRT and latent regression model (population or conditioning model) is comparable across countries. While the reliabilities reach levels above 0.80 for Mathematics, Reading, and Science in all countries except for Senegal, it is important to keep in mind that this is not to be confused with a classical reliability coefficient, as it is based on more than the item responses. Comparisons among individual students are not appropriate because the apparent accuracy of the measures is obtained by statistically adjusting the estimates based on background data. This approach does provide improved behavior of subgroup estimates, even if the plausible values obtained using this methodology are not suitable for comparisons of individuals (e.g., Mislevy and Sheehan, 1987; von Davier, Sinharay, Oranje, and Beaton, 2006).

### TRANSFORMING THE PLAUSIBLE VALUES TO PISA SCALES

By using common items between PISA-D and PISA, and by applying the same population modeling methods to generate the plausible values based on those linked scales, the same linear transformation can be applied to transform PISA-D plausible value to the PISA reporting scale. Table 12.5 provides these transformation coefficients. The coefficient *A* adjusts for the variability (standard deviation) in the plausible values, while coefficient *B* coefficient adjusts for the scale location (mean).

<sup>4</sup> Reading Components items are included in this scale.

Table 12.5 PISA 2015 transformation coefficients

Domain	A	B
Mathematics	135.9030	514.1848
Reading	131.5806	437.9583
Science	168.3189	494.5360

Table 12.6 shows the average transformed plausible values for each cognitive domain by country as well as the resampling-based standard errors.

Table 12.6 Average plausible values (PVs) and resampling-based standard errors (SE) by country for the PISA domains of Mathematics, Reading, and Science

Country	Mathematics		Reading		Science	
	Average PV	SE	Average PV	SE	Average PV	SE
<i>PISA-D International Average</i>	324	1.19	346	1.21	349	1.04
Cambodia	325	2.67	321	2.07	330	1.91
Ecuador	377	3.05	409	3.40	399	2.94
Guatemala	334	3.25	369	3.50	365	2.92
Honduras	343	3.49	371	3.46	370	2.86
Paraguay	326	2.91	370	3.68	358	3.32
Senegal	304	2.63	306	1.78	309	1.82
Zambia	258	3.88	275	3.85	309	3.05

## INTERNATIONAL CHARACTERISTICS OF THE ITEM POOL

This section provides an overview of test targeting and the domain inter-correlations.

### Test targeting

In addition to identifying the relative discrimination and difficulty of items, IRT can be used to summarise the characteristics of various subpopulations of students.

#### *Test characteristic curve (TCC) and test information function (TIF)*

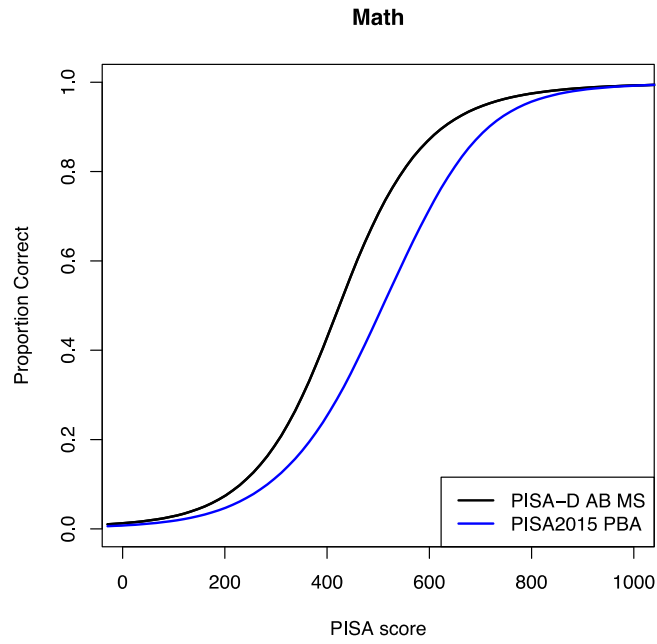
Based on the item parameters obtained from the IRT scaling, test characteristic curves (TCCs) were generated for each cognitive domain. Each domain TCC was obtained by adding all item characteristic curves (ICCs) pertinent to that domain. These TCCs (Figures 12.5 to 12.7) show the targeting of the average test difficulty across all forms on the PISA scale. This depiction is also useful for determining how students who took the PISA-D Main Survey would perform on the PISA scale. Note that only full credit scores were taken into account when probabilities were calculated, and partial credit scores were not considered for the polytomous items. This might have moved the resulting TIFs and TCCs slightly to the right; that is, the actual targeted student population on the PISA-D may be lower than presented in the figures when the partial credit scores are included.



Figures 12.5 to 12.7 also illustrate how the overall PISA-D Main Survey TCCs (black line) compare to the ones based on the paper-based (PBA) items from the PISA 2015 Main Survey (blue line) in terms of the expected proportion correct on y-axis. In Mathematics and Science domains, it is clear that PISA-D TCCs show a steeper curve at the lower end compared to PBA items from the PISA 2015 Main Survey. On the other hand, in the Reading domain, the two curves for PISA-D and PISA 2015 PBA Main Survey are very close, but the TCC of Reading Components items (red line) shows a steeper shape at the lower end of the PISA scale compared to Reading items. This shows that Reading Components items target students at the lower end of the proficiency scale compared to the Reading items overall.

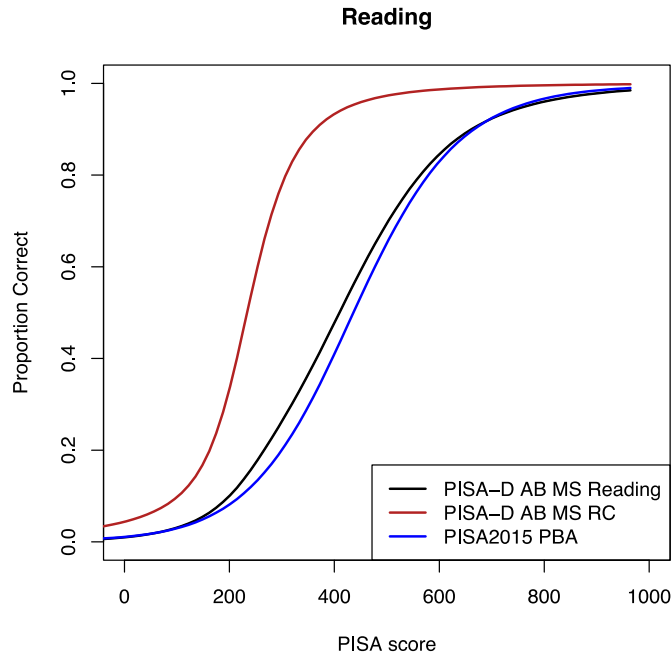
■ Figure 12.5 ■

**Test characteristic curve of PISA-D compared to the PISA 2015 Main Survey PBA items in Math**



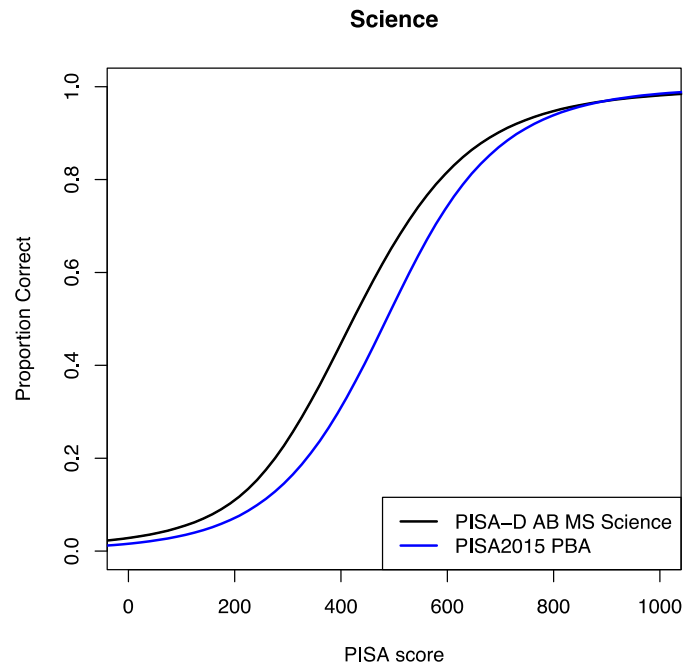
■ Figure 12.6 ■

**Test characteristic curve of PISA-D compared to the PISA 2015 Main Survey PBA items in Reading**



■ Figure 12.7 ■

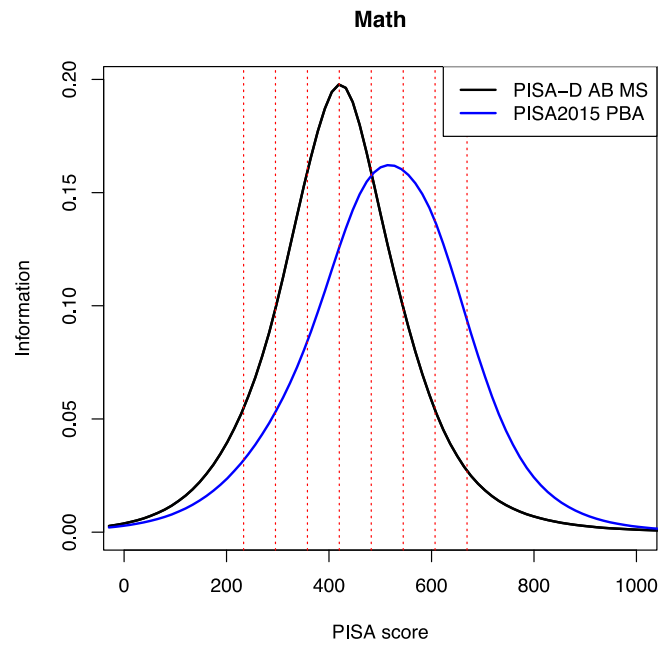
**Test characteristic curve of PISA-D compared to the PISA 2015 Main Survey PBA items in Science**



Test information functions (TIFs) for each domain were also generated based on the estimated item parameters (discrimination and difficulty) in order to examine the measurement accuracy provided by the PISA-D Main Survey assessments. TIFs are useful for examining how measurement is targeted on the PISA (and PISA-D) scales and if it is appropriate for measuring the targeted populations. In Figures 12.8 to 12.10, the highest point of the curve shows where measurement is most accurate. In each figure, the black line (labeled PISA-D AB) represents the expected test information within a one-hour assessment period (i.e., averaged across four clusters and multiplied by 2 to reflect that students took two clusters in one hour). Orange dotted vertical lines represent cut scores for proficiency levels. As shown, the highest point of the curves occurred at a scale score of approximately 420 in Mathematics, 410 in Reading, 200 in Reading Components, and 400 in Science. This corresponds that PISA-D targets nearly 1.5 levels lower than PISA for Mathematics, and nearly 1 level lower for Science. For Reading, PISA-D targets a similar level as PISA, but Reading Components improves the measurement accuracy for lower levels. This confirms that PISA-D was able to target lower-performing students, as expected. When compared to the TIFs based on the PISA 2015 PBA items (blue line), low-performing students were successfully targeted in the Mathematics and Science domains. Reading Components items, specifically, also successfully targeted low-performing students. In Reading, the TIF appeared bimodal, one at a lower PISA scale score and another at a very similar scale score as in the PISA 2015 PBA Main Survey. This shows that the target population in PISA-D was not that different from that of the PISA 2015 PBA Main Survey.

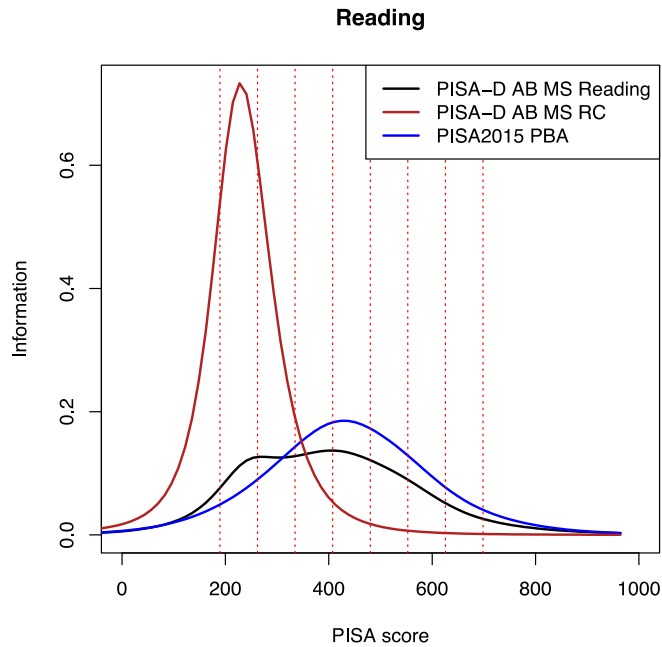
■ Figure 12.8 ■

**Test information function of PISA-D compared to the PISA 2015 Main Survey PBA items in Math**



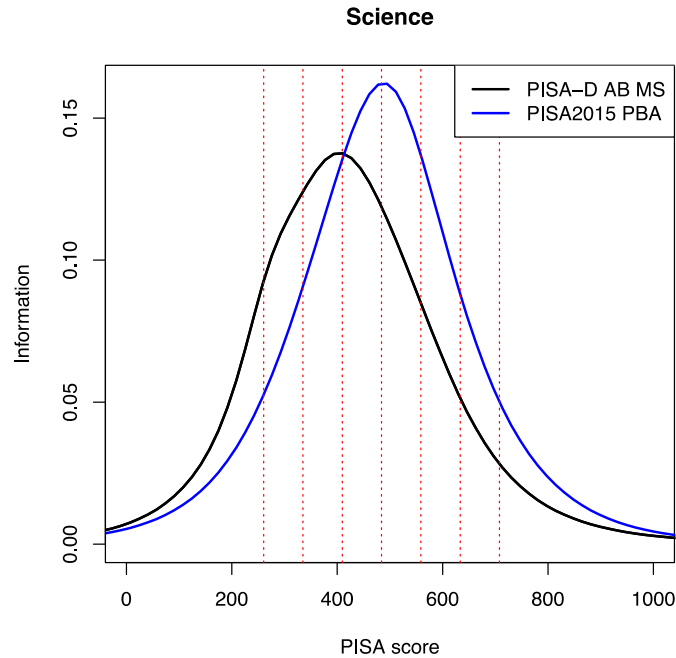
■ Figure 12.9 ■

**Test Information function of PISA-D compared to the PISA 2015 Main Survey PBA items in Reading**



■ Figure 12.10 ■

**Test information function of PISA-D compared to the PISA 2015 Main Survey PBA items in Science**



*Item RP62 values*

A specific value—the response probability (RP)—can be assigned to each item on a scale according to its discrimination and difficulty, similar to how students receive specific scaled scores according to their performance on assessment items (Kirsch, de Jong, Lafontaine, McQueen, Mendelovits, and Monseur, 2002). Chapter 15 describes how items can be placed along a scale based on RP values and how these values can be used to describe different proficiency levels.

After the estimation of item parameters in the item calibration stage, RP values were calculated for each item, and then items were classified into proficiency levels within the cognitive domain. Likewise, after generation of the plausible values, respondents can be classified into proficiency levels for each cognitive domain. The purpose of classifying items and respondents into levels is to provide more descriptive information about group proficiencies. The different item levels provide information about the underlying characteristics of an item as it relates to the domain (such as item difficulty); the higher the difficulty, the higher the level. Similar to PISA, in PISA-D, an RP62 value was used to classify items into levels. Respondents with a proficiency located below this point have a lower probability than the chosen RP62 value, and respondents with a proficiency above this point have a higher probability (that is  $> 0.62$ ) of solving the item. The RP62 values for all items are presented in Annex A together with the final item parameters obtained from the IRT scaling. The classification of respondents into different levels is done by transforming plausible values into PISA scale scores. Each level is defined by certain score boundaries for each cognitive domain. Tables 12.7 to 12.9 show the overall score boundaries

used for each cognitive domain along with the percentage of items and proportion of respondents classified at each level of proficiency. In PISA-D, the lowest level (1c) has been added for each domain to describe the lower end of the proficiency distribution. The choice of score boundaries for the lowest levels is explained in Chapter 15; apart from level 1c, all other levels have been defined in previous PISA cycles. In addition, the classification of PISA 2015 PBA items is given in each table for comparison. In each cognitive domain, the majority of PISA-D Main Survey items were classified as Level 3 or below, while at least a plurality of PISA 2015 PBA items were classified as Level 3 or Level 4.

Table 12.7 **Item and respondent classification for each score boundary in Math**

Level	Score points on the PISA scale	Number of items	Percentage of items	PISA 2015 MS PBA (%) <sup>5</sup>
6	Higher than 669.30	1	1.6	13.1
5	Higher than 606.99 and less than or equal to 669.30	4	6.5	10.7
4	Higher than 544.68 and less than or equal to 606.99	7	11.3	23.8
3	Higher than 482.38 and less than or equal to 544.68	9	14.5	22.1
2	Higher than 420.07 and less than or equal to 482.38	25	40.3	17.2
1a	Higher than 357.77 and less than or equal to 420.07	8	12.9	7.4
1b	Higher than 295.47 and less than or equal to 357.77	6	9.7	5.7
1c	Higher than 233.17 and less than or equal to 295.47	2	3.2	
Below 1c	Less than 233.17	0	0.0	
Total		62	100	100

<sup>5</sup> Includes PISA 2015 Main Survey PBA items and historic PBA items.

Table 12.8 Item and respondent classification for each score boundary in Reading

Level	Score points on the PISA scale	Number of items		Percentage of items		PISA 2015 MS PBA (%)
		Reading	RC	Reading	RC	
6	Higher than 698.32	1	0	1.5	0.0	3.8
5	Higher than 625.61 and less than or equal to 698.32	2	0	3.0	0.0	4.5
4	Higher than 552.89 and less than or equal to 625.61	13	0	19.7	0.0	15.8
3	Higher than 480.18 and less than or equal to 552.89	10	0	15.2	0.0	21.1
2	Higher than 407.47 and less than or equal to 480.18	14	1	21.2	1.3	24.8
1a	Higher than 334.75 and less than or equal to 407.47	18	2	27.3	2.5	23.3
1b	Higher than 262.04 and less than or equal to 334.75	3	30	4.6	38.0	5.3
1c	Higher than 189.33 and less than or equal to 262.04	5	46	7.6	58.2	1.5
Below 1c	Less than 189.33	0	0	0.0	0.00	
Total		66	79	100	100	100

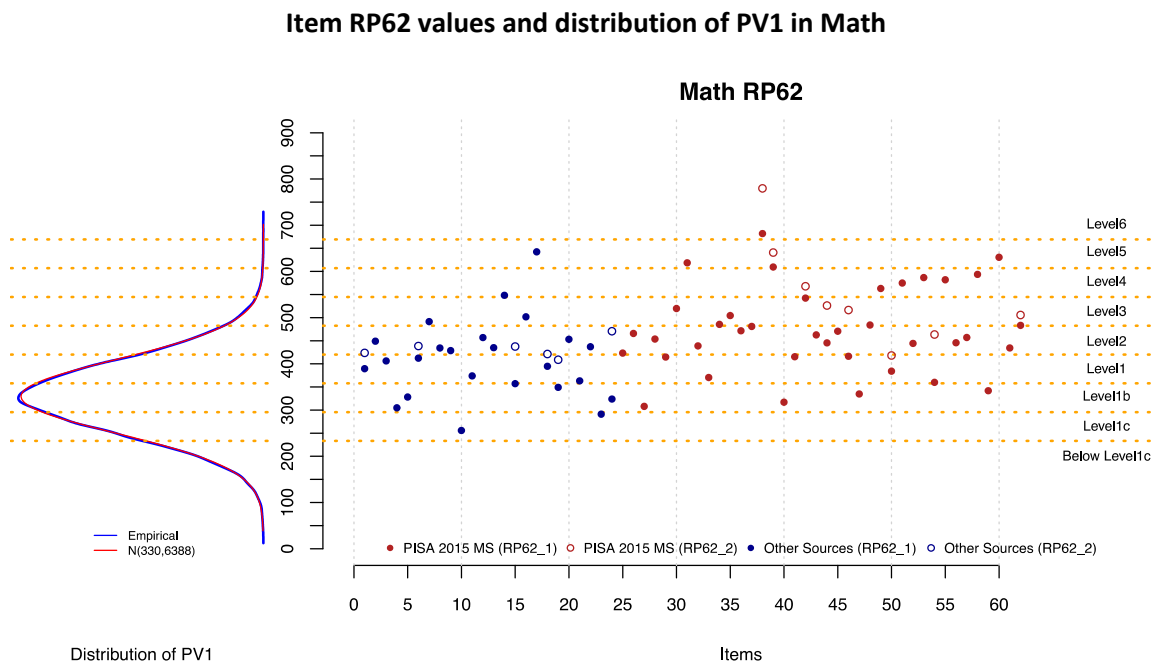
Table 12.9 Item and respondent classification for each score boundary in Science

Level	Score points on the PISA scale	Number of items	Percentage of items	PISA 2015 MS PBA (%) <sup>6</sup>
6	Higher than 707.93	2	3.0	4.6
5	Higher than 633.33 and less than or equal to 707.93	2	3.0	7.4
4	Higher than 558.73 and less than or equal to 633.33	12	18.2	25.0
3	Higher than 484.14 and less than or equal to 558.73	17	25.8	36.1
2	Higher than 409.54 and less than or equal to 484.14	20	30.3	21.3
1a	Higher than 334.94 and less than or equal to 409.54	10	15.2	4.6
1b	Higher than 260.54 and less than or equal to 334.94	3	4.6	0.9
Below 1b	Less than 260.54	0	0.0	0.0
Total		66	100	100

<sup>6</sup> Includes PISA 2015 Main Survey PBA items and historic PBA items.

Because RP62 values and the transformed plausible values are on the same PISA scale, the distribution of respondents' latent abilities and item RP62 values can be represented on the common scale. Figures 12.11 to 12.13 illustrate the distribution of the first plausible value (PV1) along with item RP62 values on the PISA scale separately for each cognitive domain for the PISA-D Main Survey data. Dichotomous items have only a solid dot, and polytomous items have both a solid dot (indicating the first threshold from 0 to 1) and a hollow dot (indicating the second threshold from 1 to 2) for each item. In each domain, different colors indicate the source of the items, whether it was a PISA (red) or non-PISA item (blue). On the left side, the distribution of plausible values is plotted. In each figure, the blue line indicates the empirical density of the plausible values across countries, and the red line indicates the theoretical normal distribution with mean of plausible values and the variance of plausible values in each domain across countries. Specifically,  $N(330, 79.93^2)$  for Mathematics,  $N(351, 84.41^2)$  for Reading, and  $N(352, 68.60^2)$  for Science are displayed as red lines.

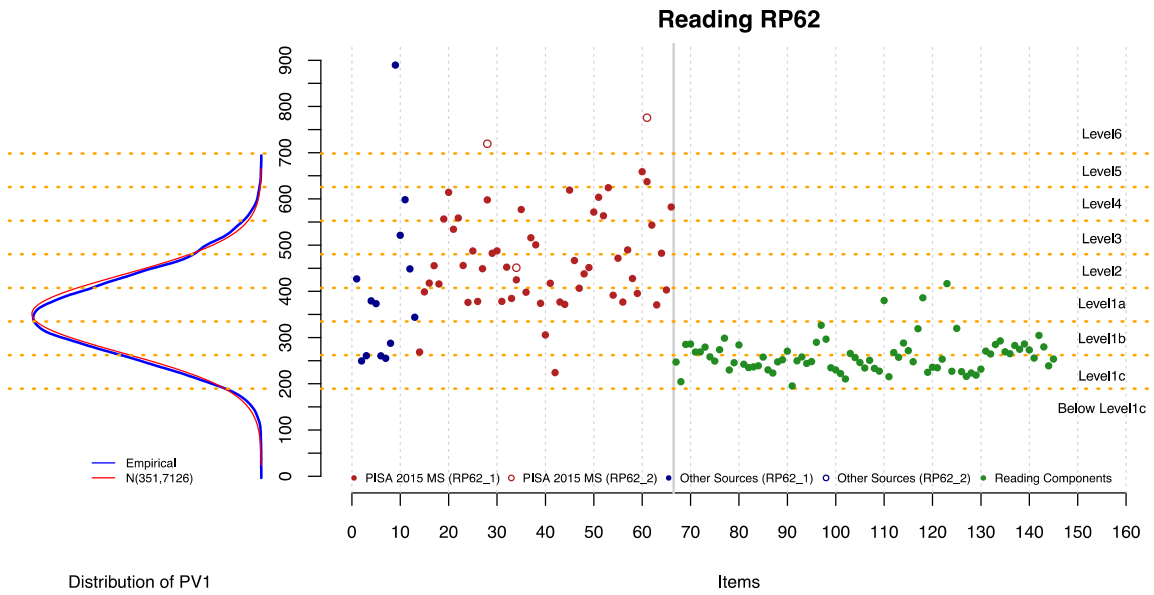
■ Figure 12.11 ■





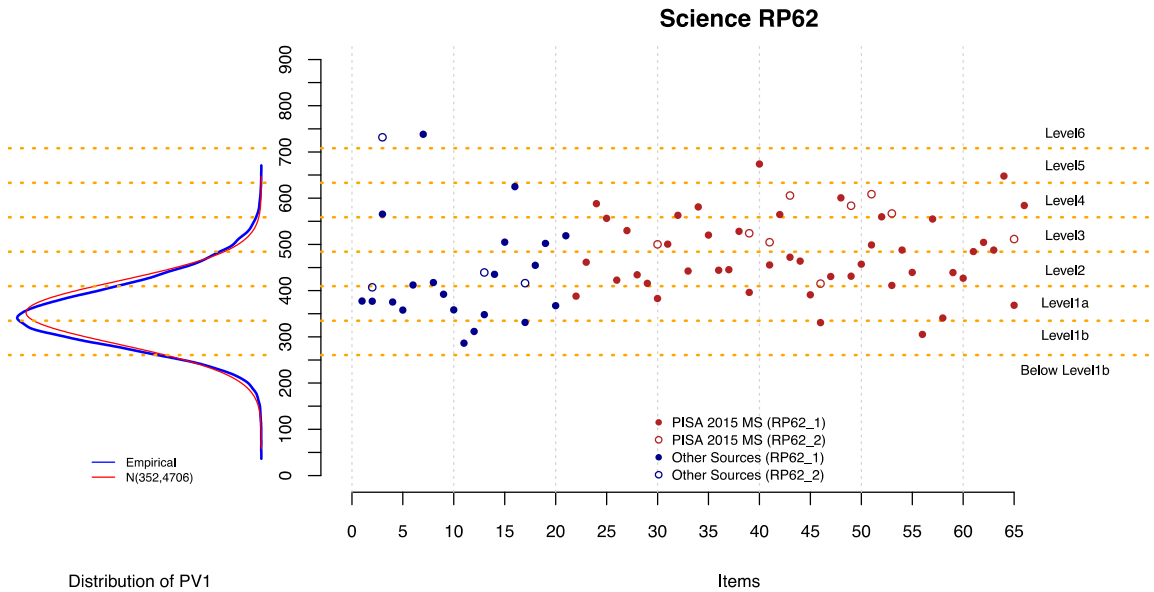
■ Figure 12.12 ■

**Item RP62 values and distribution of PV1 in Reading (and Reading Components)**



■ Figure 12.13 ■

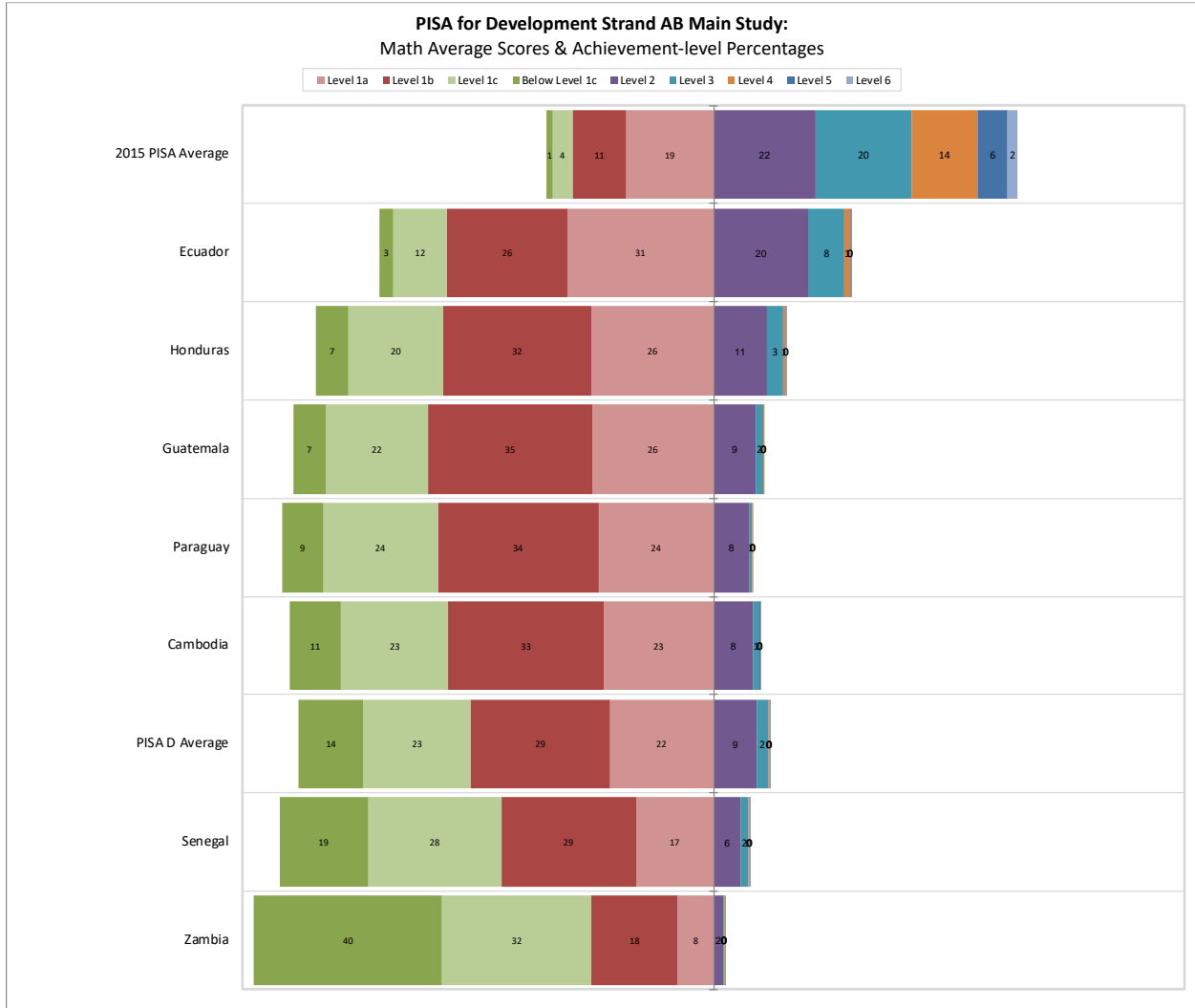
**Item RP62 values and distribution of PV1 in Science**



Figures 12.14 to 12.16 show the percentage of respondents per country at each level of proficiency for each cognitive domain, using Level 2 as a reference. The PISA 2015 Main Survey average and the PISA-D average are illustrated together to better illustrate the comparison.

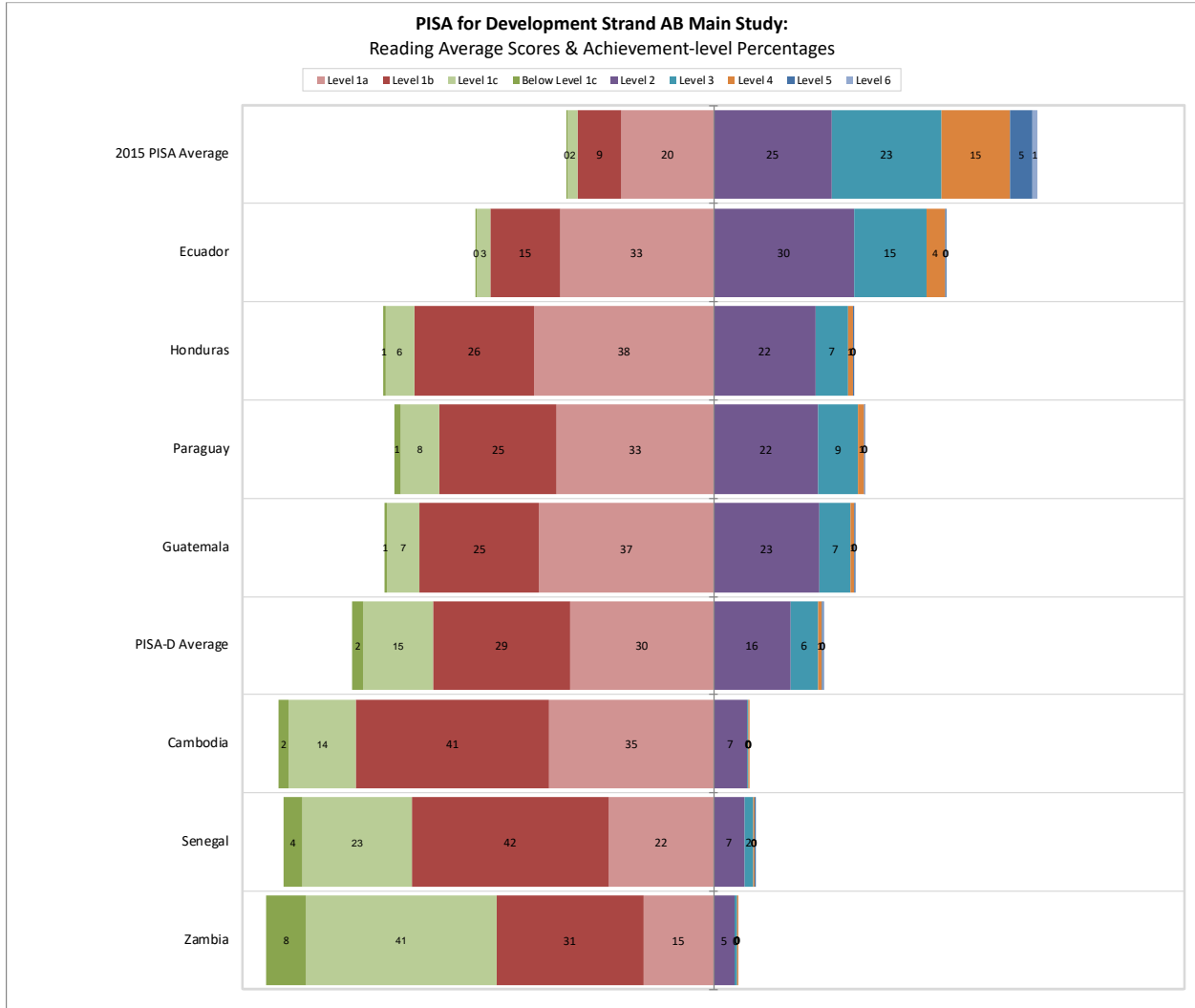
■ Figure 12.14 ■

Percentage of respondents per country at each level of proficiency for Mathematics



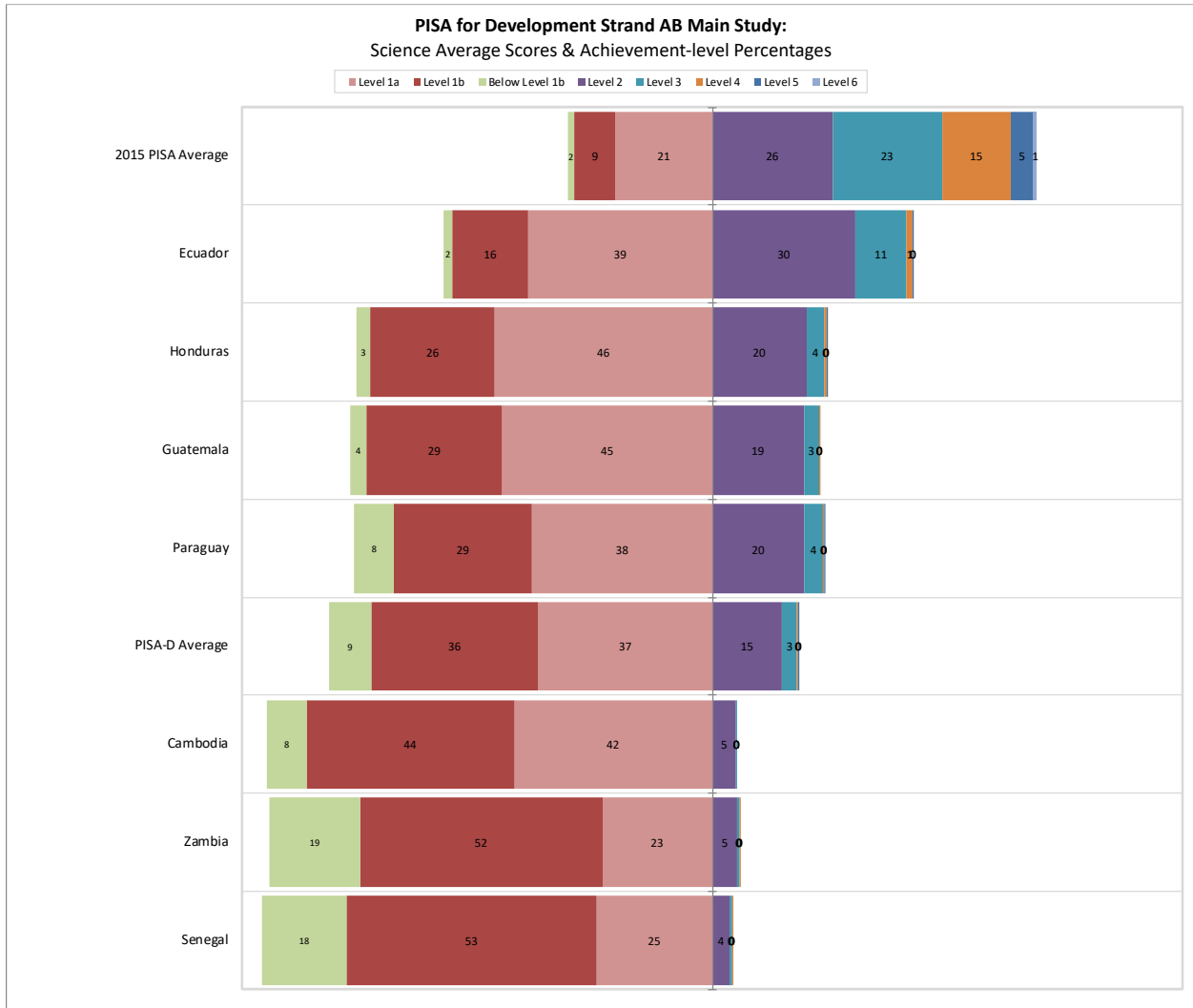
■ Figure 12.15 ■

Percentage of respondents per country at each level of proficiency for Reading



■ Figure 12.16 ■

**Percentage of respondents per country at each level of proficiency for Science**



**Domain inter-correlations**

Estimated correlations between the PISA-D domains, based on the 10 plausible values and averaged across all countries, are presented in Table 12.10. Overall, the correlations are quite high, as expected, yet there is still some separation between each of the domains. The estimated correlations at the national level are presented in Table 12.11.

Table 12.10 Domain inter-correlations

Domain		Reading	Science
Math	Average	0.77	0.81
	Range	0.45 - 0.86	0.63 - 0.88
Reading	Average	---	0.82
	Range	---	0.64 - 0.88

Table 12.11 National-level domain inter-correlations based on 10 PVs

Countries	Math & Reading	Math & Science	Reading & Science
Cambodia	0.79	0.76	0.76
Ecuador	0.80	0.84	0.86
Guatemala	0.86	0.87	0.88
Honduras	0.84	0.88	0.85
Paraguay	0.82	0.84	0.87
Senegal	0.45	0.63	0.64

### References

**Efron, B.** (1982), "The jackknife, the bootstrap, and other resampling plans", *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, Vol. 38.

**Kirsch I., J. de Jong, D. Lafontaine, J. McQueen, J. Mendelovits and C. Monseur** (2002), *Reading for Change – Performance and Engagement Across Countries*, retrieved from OECD website: <http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpi/sa/33690904.pdf>.

**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.

**von Davier, M., S. Sinharay, A. Oranje and A. Beaton** (2006), "The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions", In C. R. Rao and S. Sinharay (ed.), *Handbook of Statistics* (Vol. 26, pp. 1039–1055), Elsevier, Amsterdam, Netherlands.