# Chapter 13: SCALING AND POPULATION MODELLING OUTCOMES FOR COGNITIVE DATA

## INTRODUCTION

This chapter presents the outcomes of applying the item response theory (IRT) scaling and population models to the PISA-D Strand C cognitive data. Outcomes include the percent of common and unique item parameters, percent of respondents in each plausible level, classification of items using RP62 values, the test characteristic curve and the test information function for each domain.

## RESULTS OF IRT SCALING AND POPULATION MODELLING

### Scaling outcomes

As elaborated in Chapter 10, at the beginning of the scaling process, all the item parameters (including the group-specific item parameters) were fixed to the parameters that had been estimated in the PISA-D Strand A/B Main Survey, many of which had been fixed to the PISA 2015 parameters, as applicable. This method provided a strong linkage across the different assessments, including PISA-D Strand C, PISA-D Strand A/B and PISA. When the item parameters obtained from PISA-D Strand A/B did not fit the data for PISA-D Strand C in specific country-by-language groups, new item parameters were allowed to be estimated for the group exhibiting misfit, either together with another group or by itself. When the level of misfit was large (RMSD > 0.4) or when the slope parameter was close to 0 or negative, the item was excluded from the scaling for the group.

Table 13.1 presents the percentage of common and unique item-by-group parameters for each domain, and Annex A presents the international item parameters for each item. In total, 92% of the item-by-group parameters for Math, 67% for Reading and 73% for Reading Components were in common with PISA Strands A/B, supporting a strong link between PISA-D Strand C and PISA-D Strand A/B. In addition, 38% of the item-by-group parameters for Math and 50% for Reading were the same as in PISA 2015, supporting a stable link between PISA-D Strand C and PISA 2015. There were no Reading Component items administered in PISA 2015, and therefore, these are reported in a separate row in the table below. The item-by-group parameters that are different from Strand A/B (7% for Math, 30% for Reading and 25% for Reading Components) do not contribute to the linking between the two assessments, but instead reduce the measurement error within each country-by-language group. In general, the large amount of commonality of the item-by-group parameters support that a strong link was established across PISA-D Strand C, PISA-D Strand A/B and PISA 2015.

**Table 13.1 Percentage of common and unique item-by-group parameters in each domain**

| Item-by-group parameters | Math | Reading | Reading Components |
|---|---|---|---|
| Same as Strand A/B international parameters and same as PISA 2015 int'l parameters | 38 | 50 | -- |
| Same as Strand A/B international parameters but different from PISA 2015 int'l parameters | 54 | 17 | -- |
| Same as Strand A/B international parameters, item not included in PISA 2015 | -- | -- | 73 |
| Different from Strand A/B international parameters but common for two or more groups | 1 | 5 | 7 |
| Different from Strand A/B international parameters and unique to a specific group | 6 | 25 | 18 |
| Deleted | 1 | 4 | 3 |
| **Total (%)** | **100** | **100** | **100** |
| **Number of items** | **35** | **22** | **50** |

Tables 13.2 to 13.4 present the percentage of common and unique item parameters for each domain, disaggregated by country-by-language group.

**Table 13.2 Percentage of common and unique item parameters for Math**

| Item parameters | Guatemala | Honduras | Panama | Paraguay | Senegal-French | Senegal-Wolof |
|---|---|---|---|---|---|---|
| Same as Strand A/B int'l parameters and same as PISA 2015 int'l parameters | 37 | 43 | 40 | 37 | 37 | 34 |
| Same as Strand A/B int'l parameters but different from PISA 2015 int'l parameters | 54 | 57 | 57 | 49 | 51 | 54 |
| Different from Strand A/B int'l parameters but common for two or more groups | 3 | -- | -- | 3 | -- | -- |
| Different from Strand A/B int'l parameters and unique to a specific group | 3 | -- | 3 | 9 | 11 | 11 |
| Deleted | 3 | -- | -- | 3 | -- | -- |
| **Total (%)** | **100** | **100** | **100** | **100** | **100** | **100** |

**Table 13.3 Percentage of common and unique item parameters for Reading**

| Item parameters | Guatemala | Honduras | Panama | Paraguay | Senegal-French | Senegal-Wolof |
|---|---|---|---|---|---|---|
| Same as Strand A/B int'l parameters and same as PISA 2015 int'l parameters | 50 | 59 | 50 | 41 | 50 | 50 |
| Same as Strand A/B int'l parameters but different from PISA 2015 int'l parameters | 23 | 23 | 5 | 14 | 14 | 23 |
| Different from Strand A/B int'l parameters but common for two or more groups | 9 | -- | 5 | 9 | 5 | -- |
| Different from Strand A/B int'l parameters and unique to a specific group | 18 | 14 | 41 | 32 | 27 | 18 |
| Deleted | -- | 5 | -- | 5 | 5 | 9 |
| **Total (%)** | **100** | **100** | **100** | **100** | **100** | **100** |

**Table 13.4 Percentage of common and unique item parameters for Reading Components**

| Item parameters | Guatemala | Honduras | Panama | Paraguay | Senegal-French | Senegal-Wolof |
|---|---|---|---|---|---|---|
| Same as Strand A/B int'l parameters and same as PISA 2015 int'l parameters | -- | -- | -- | -- | -- | -- |
| Same as Strand A/B int'l parameters but different from PISA 2015 int'l parameters | -- | -- | -- | -- | -- | -- |
| Same as Strand A/B int'l parameters, item not included in PISA 2015 | 92 | 80 | 76 | 80 | 65 | 43 |
| Different from Strand A/B int'l parameters but common for two or more groups | 2 | 10 | 10 | 6 | 8 | 4 |
| Different from Strand A/B int'l parameters and unique to a specific group | 4 | 8 | 12 | 12 | 25 | 47 |
| Deleted | 2 | 2 | 2 | 2 | 2 | 6 |
| **Total (%)** | **100** | **100** | **100** | **100** | **100** | **100** |

### Plausible levels

As explained in Chapter 10, proficiency results for PISA-D Strand C are reported as *plausible levels* instead of as plausible values. This is due to data quality issues and the relatively high level of

measurement uncertainty. Plausible levels are estimates of the proficiency level of the individuals based on their performance in the assessment. Table 13.5 shows the percentage of respondents in each proficiency level for Math for each participating country, and Table 13.6 shows the results for Reading.

**Table 13.5 Percentage of respondents in each proficiency level for Math**

| Level | Guatemala | Honduras | Panama | Paraguay | Senegal |
|---|---|---|---|---|---|
| Level 2 and above | -- | 3 | 2 | -- | -- |
| Level 1a | 4 | 15 | 7 | 2 | 3 |
| Level 1b | 11 | 33 | 19 | 7 | 20 |
| Level 1c | 23 | 31 | 25 | 17 | 39 |
| Below Level 1c | 63 | 18 | 47 | 74 | 38 |
| **Total (%)** | **100** | **100** | **100** | **100** | **100** |

**Table 13.6 Percentage of respondents in each proficiency level for Reading**

| Level | Guatemala | Honduras | Panama | Paraguay | Senegal |
|---|---|---|---|---|---|
| Level 2 and above | -- | 2 | 4 | 1 | -- |
| Level 1a | 8 | 20 | 21 | 5 | 3 |
| Level 1b | 37 | 46 | 35 | 33 | 31 |
| Level 1c | 44 | 29 | 36 | 47 | 57 |
| Below Level 1c | 11 | 3 | 4 | 14 | 9 |
| **Total (%)** | **100** | **100** | **100** | **100** | **100** |

**Reliability of the plausible levels**

The reliability of the plausible levels was estimated using the commonly used formula: *1 – (expected error variance/total variance)*, similar to the methodology used in PISA 2018 (OECD, 2020). However, note that the reported ordinal plausible levels were used for calculating the reliability instead of the continuous plausible values. The expected error variance was the weighted average of the imputation variance (i.e. the variance across the 10 plausible levels, which is an expression of the posterior measurement error). The total variance was estimated using the weighted variance for all students, applying the senate sampling weights.

The reliability of the plausible levels is presented in Table 13.7 for Math and in Table 13.8 for Reading. Note that the literacy-related non-respondents (LRNR) as well as those that did not respond to any of the cognitive items were automatically assigned the lowest plausible level (i.e. below level 1c) for all 10 plausible levels, for both Math and Reading. Therefore, two types

of reliabilities are reported below. In the first column, labelled "All", we present the reliability calculated using all respondents who received plausible levels, while in the second column, labelled "Normal", we present the reliability excluding respondents who automatically received the lowest plausible level. The last column in the table, labelled "% Normal", presents the proportion of respondents in each country that were used in this last calculation. It is important to keep in mind that the reliabilities based on all respondents are slightly inflated because of the cases that were automatically assigned the lowest plausible level for all 10 of their plausible levels in both Math and Reading. Consequently, the lower the percent of normal cases within a country, the bigger the difference can be between the reliability based on all respondents and the reliability based only on the normal cases.

**Table 13.7 Reliability of the plausible levels for Math**

| Country | All | Normal | % Normal |
|---|---|---|---|
| Guatemala | 0.826 | 0.811 | 83.7 |
| Honduras | 0.787 | 0.763 | 95.0 |
| Panama | 0.812 | 0.812 | 100.0 |
| Paraguay | 0.718 | 0.704 | 85.2 |
| Senegal | 0.709 | 0.690 | 95.0 |

**Table 13.8 Reliability of the plausible levels for Reading**

| Country | All | Normal | % Normal |
|---|---|---|---|
| Guatemala | 0.871 | 0.772 | 83.7 |
| Honduras | 0.830 | 0.783 | 95.0 |
| Panama | 0.838 | 0.838 | 100.0 |
| Paraguay | 0.859 | 0.768 | 85.2 |
| Senegal | 0.765 | 0.708 | 95.0 |

## CHARACTERISTICS OF THE ITEMS AND OVERALL ASSESSMENT

### Item RP62 values

After estimating the item parameters in the item calibration stage, response probability 62 (RP62) values were calculated for each item. RP62 values provide information on the difficulty of the item – respondents with a proficiency below the RP62 value of an item have less than a 62% probability of responding to the item correctly, while respondents with a proficiency above the RP62 value of an item have more than a 62% probability of responding to the item correctly. Thus, taking both item slope and difficulty into account, the more difficult an item, the higher the RP62 value will be (Kirsch, de Jong, Lafontaine, McQueen, Mendelovits, & Monseur, 2002).

Subsequently, using the domain-specific cut-off scores on the PISA scale and the RP62 value for each item, the items were classified into plausible levels. For polytomous items, only the second RP62 value (associated with obtaining the highest score on the item) was used for the classification of the item. Table 13.9 and Table 13.10 show the number and percentage of items that were classified into each proficiency level for Math and Reading, respectively, and Annex A presents the RP62 value for each item. For Math, 29% of the items were classified below Level 2, as were 64% of the Reading items. All Reading Components items were classified below Level 2.

**Table 13.9 Item classification using RP62 values for Math**

| Level | Score points on the PISA scale | # of items | % of items |
|---|---|---|---|
| Level 2 and above | 420.07 or above | 25 | 71.4 |
| Level 1a | 357.77 or above | 7 | 20.0 |
| Level 1b | 295.47 or above | 4 | 11.4 |
| Level 1c | 233.17 or above | 2 | 5.7 |
| Below Level 1c | Below 233.17 | 0 | 0.0 |
| **Total** | | **35** | **100.0** |

**Table 13.10 Item classification using RP62 values for Reading**

| Level | Score points on the PISA scale | Reading | | Reading Components | |
|---|---|---|---|---|---|
| | | # of items | % of items | # of items | % of items |
| Level 2 and above | 407.47 or above | 8 | 36.4 | 0 | 0.0 |
| Level 1a | 334.75 or above | 9 | 40.9 | 1 | 2.0 |
| Level 1b | 262.04 or above | 2 | 9.1 | 14 | 28.0 |
| Level 1c | 189.33 or above | 3 | 13.6 | 34 | 68.0 |
| Below Level 1c | Below 189.33 | 0 | 0.0 | 1 | 2.0 |
| **Total** | | **22** | **100.0** | **50** | **100.0** |

### Test targeting

For each cognitive domain, the test characteristic curve (TCC) was generated by adding all the IRT-based item characteristic curves (ICCs) of the items included in the domain. The TCCs show the proficiency level of the population that was best targeted across all forms in the assessment, which is useful for determining how students who took the PISA-D Strand C Main Survey would perform on the PISA scale.

The TCC for math is presented in Figure 13.1 while the TCC for Reading is presented in Figure 13.2. For reference, the TCCs for PISA-D Strand A/B and PISA 2015 PBA items[1] are also presented in the figures. For both Math and Reading, and especially Reading Components, it is clear that

PISA-D Strand C targeted respondents that were less proficient than those targeted by PISA-D Strand A/B and PISA 2015 PBA. Note that for polytomous items, only full credit scores were taken into account, following the method used in PISA-D Strand A/B (OECD, 2019), which might have shifted the resulting TCCs slightly to the right (i.e. when the partial credit scores are included, the actual population targeted by PISA-D Strand C may be lower than what is presented in the figures).

**Figure 13.1 Test characteristic curve for Math**

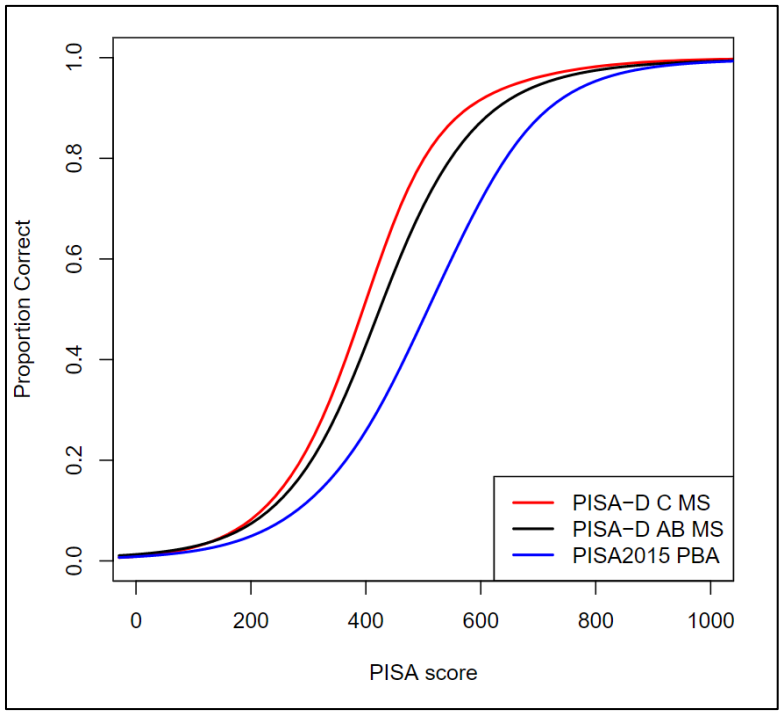
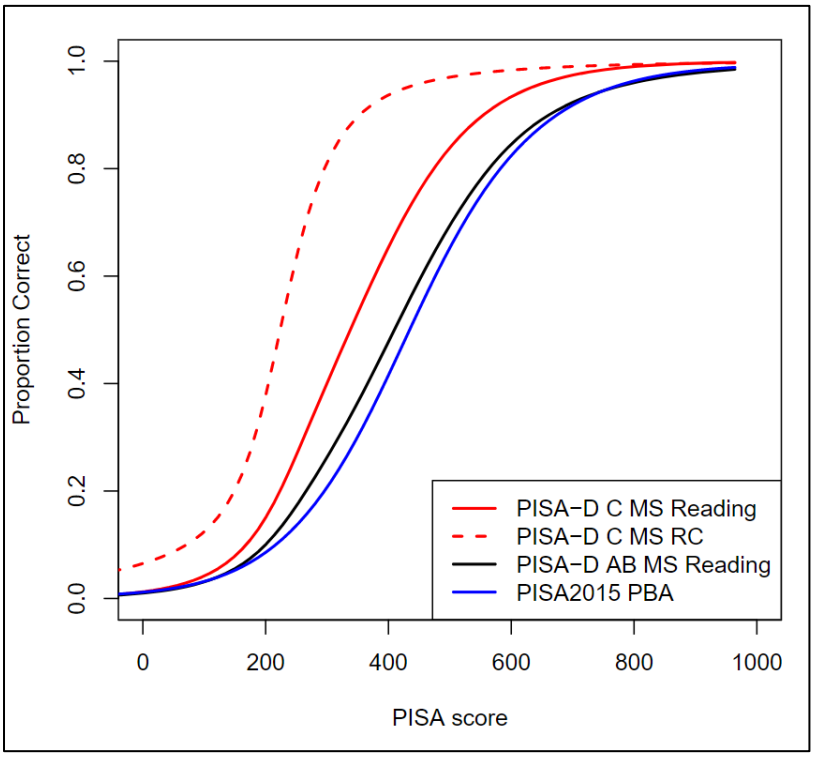


**Figure 13.2 Test characteristic curve for Reading and Reading Commponents**




In order to examine measurement accuracy, test information functions (TIFs) were generated based on the final international item parameters. Note that the score at which the curve peaks is where measurement is the most accurate. TIFs are useful for examining how measurement is targeted in the PISA-D Strand C Main Survey and whether the assessment is appropriate for measuring the targeted populations.

According to the PISA-D Strand C Main Survey design, respondents first took the Core Module which consisted of five Math and five Reading items. If respondents provided fewer than two correct responses to the 10 items in the Core Module, they were considered to have failed the Core Module. These respondents subsequently took all four Reading Components clusters, including one Sentence Processing cluster (24 items) and three Passage Comprehension clusters (total of 26 items), but did not take any Math or Reading clusters. On the other hand, the respondents who provided at least two correct responses to the 10 items in the Core Module were considered to have passed the Core Module. These respondents subsequently took the Reading Components Sentence Processing cluster (24 items), one block of the Reading Components Passage Comprehension cluster (seven to 10 items), one or two blocks of Math (10 items per block), and one or two blocks of Reading (5 to 6 items per block).

Figure 13.3 presents the TIF for Math for respondents who passed the Core Module, based on the average number of Math items across the different forms. Specifically, the solid red curve represents the TIF for the Math items included in the Core Module and two Math clusters (average of 25 items), while the dotted red curve represents the TIF for the Math items included

in the Core Module and only one Math cluster (average of 15 items). For reference, the TIF for PISA-D Strand A/B is presented as a black curve (average of approximately 31 items), while the TIF for PISA 2015 PBA is presented as a blue curve (average of approximately 24 items). The figure shows that for Math, measurement accuracy for PISA-D Strand C is the highest at approximately 380 points, which is slightly lower than the score at which measurement accuracy peaked for PISA-D Strand A/B (at 420 points), and more than 150 points lower than the score at which measurement accuracy peaked for PISA 2015 PBA (at 530 points). Note that only full credit scores were taken into account for polytomous items, which might have shifted the resulting TIF slightly to the right. If the partial credit scores had been considered, the TIF might have peaked at a lower score than what is presented in the figure.

**Figure 13.3 Test information function for Math**

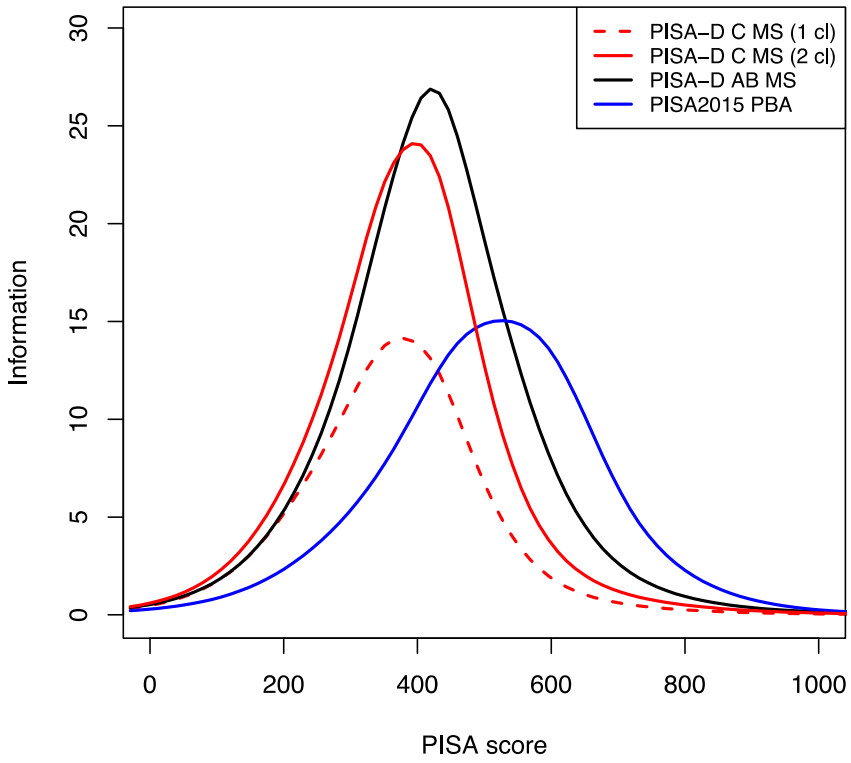
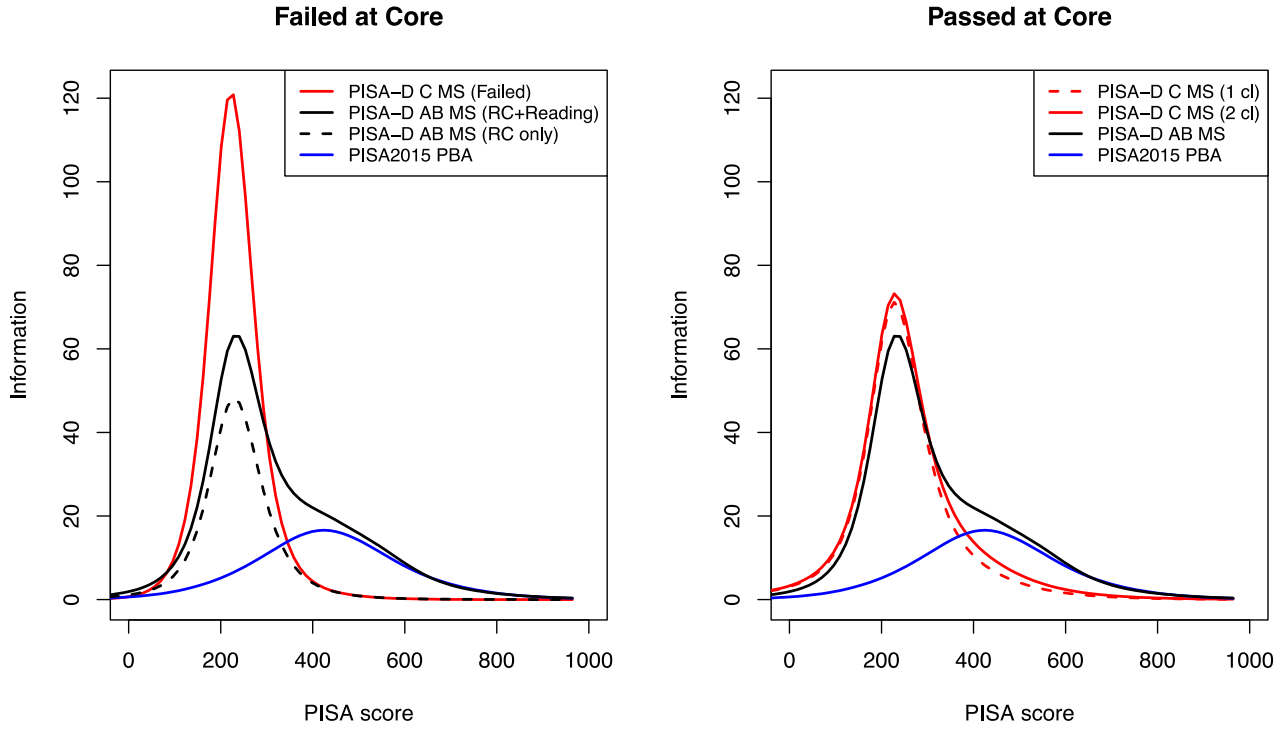



Figure 13.4 presents the TIFs for Reading for those who failed the Core Module (on the left) and those who passed it (on the right). Again, it is based on the average number of Reading items across the different forms. The solid red curve in the left panel represents the TIF for the Reading items taken by the respondents who failed the Core Module in PISA-D Strand C. As explained above, these respondents took five Reading items in the Core Module and all Reading Components items (50 items). For reference, the solid black curve presents the TIF for the Reading and Reading Components items in PISA-D Strand A/B (average of approximately 53 items), the dotted black curve presents the TIF for only the Reading Components items in PISA-D Strand A/B (average of approximately 20 items), and the blue curve represents the TIF for the

Reading items in the PISA 2015 PBA (average of approximately 29 items). The figure shows that due to the dominant impact of the Reading Components items, measurement accuracy for both PISA-D Strand C and Strand A/B peaked at approximately 230 points, which is approximately 200 points lower than the score at which the TIF peaked for PISA 2015 PBA (at 420 points).

The panel on the right presents the TIFs for Reading for those who passed the Core Module. The solid red curve takes into account information from the five Reading items in the Core Module, the Reading Components Sentence Processing cluster (24 items), one Reading Components Passage Comprehension cluster (average of approximately nine items) and two clusters of Reading (average of 12 items). The dotted red curve is similar to the solid red curve, but it takes into account information from only one cluster of Reading (average of six items) instead of two clusters of Reading. Similar to the TIFs for those who failed the Core Module, information from the Reading Components items dominated the shape of the TIFs, and as a result, only a small difference is observed between the TIF with only one cluster and two clusters of Reading. Again, for reference, the solid black curve presents the TIF for the Reading and Reading Components items in PISA-D Strand A/B (average of approximately 53 items), while the blue curve represents the TIF for the Reading items in the PISA 2015 PBA (average of approximately 29 items). This figure shows that measurement accuracy for both PISA-D Strand C and Strand A/B peaked at approximately 230 points, which is approximately 200 points lower than the score at which the TIF peaked for PISA 2015 PBA (at 420 points).

**Figure 13.4 Test information function for Reading**



### Domain inter-correlations

Table 13.11 presents the correlations between the Math and Reading domains based on the plausible levels, after applying the senate sampling weights. Again, the second column (i.e. under All) includes all respondents who received plausible levels, the third column (i.e. under Normal) includes only respondents who provided cognitive responses (i.e. excluding respondents who automatically received the lowest plausible level due to the reasons mentioned above) and the last column presents the proportion of respondents in each country that were normal cases. When all cases were included, the domain inter-correlations ranged from 0.449 (Paraguay) to 0.666 (Guatemala); when only the normal cases were included, it ranged from 0.425 (Paraguay) to 0.683 (Guatemala).

**Table 13.11 Domain inter-correlations by country**

| Country | All | Normal | % Normal |
|---|---|---|---|
| Guatemala | 0.666 | 0.683 | 83.7 |
| Honduras | 0.651 | 0.616 | 95.0 |
| Panama | 0.590 | 0.590 | 100.0 |
| Paraguay | 0.449 | 0.425 | 85.2 |
| Senegal | 0.517 | 0.477 | 95.0 |

## NOTES

[1] For the PISA 2015 PBA items, the easier clusters in Mathematics and Reading (often called 6B) were included in the analysis instead of the standard clusters (often called as 6A).

## REFERENCES


Kirsch I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002), *Reading for change – Performance and engagement across countries,* Paris, France: OECD.

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

OECD (2020), *PISA 2018 technical report,* Paris, France: OECD.