



# PISA 2015 Technical Report



P r o g r a m m e f o r I n t e r n a t i o n a l S t u d e n t A s s e s s m e n t



# PISA 2015

# Technical Report



This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Organisation or of its member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

**Photo credits:**

- © Flying Colours Ltd/Getty Images
- © Jacobs Stock Photography/Kzenon
- © khoa vu/Flickr/Getty Images
- © Mel Curtis/Corbis
- © Shutterstock/Kzenon
- © Simon Jarratt/Corbis

© OECD 2017

---

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgement of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to [rights@oecd.org](mailto:rights@oecd.org). Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at [info@copyright.com](mailto:info@copyright.com) or the Centre français d'exploitation du droit de copie (CFC) at [contact@cfcopies.com](mailto:contact@cfcopies.com).



# Table of Contents

<b>READER'S GUIDE .....</b>	17
<b>CHAPTER 1 PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT: AN OVERVIEW .....</b> 21	
<b>Introduction.....</b>	22
<b>Participation.....</b>	23
<b>Features of PISA.....</b>	23
<b>Technical innovations in PISA 2015 .....</b>	25
<b>Managing and implementing PISA.....</b>	26
<b>PISA 2015 publications.....</b>	27
<b>CHAPTER 2 TEST DESIGN AND TEST DEVELOPMENT..... 29</b>	
<b>Introduction.....</b>	30
<b>PISA 2015 integrated design .....</b>	30
▪ Minimising the distinction between major and minor domain coverage.....	30
▪ Improving comparability and stabilising trends.....	32
▪ Goals and domain coverage.....	34
▪ Studying mode effects in PISA 2015.....	35
<b>Overview of the field trial assessment design.....</b>	35
<b>Overview of the main survey assessment design .....</b> 36	
▪ Main study form assignment for the computer-based assessment.....	40
▪ Une heure (UH) form.....	42
▪ Assessment of financial literacy.....	42
<b>The 2015 assessment frameworks.....</b>	43
▪ Science .....	43
▪ Collaborative problem solving.....	43
<b>Role of the subject matter expert groups in item development.....</b>	44
<b>PISA 2015 test development .....</b>	44
▪ Computer-based assessment: Screen design and interface .....	45
▪ Trend items .....	46
▪ New Items.....	47
▪ International test development team .....	47
<b>Field trial.....</b>	49
▪ Preparation of field trial instruments .....	49
▪ Field trial coding .....	50
<b>Main survey.....</b>	51
▪ National item review following the field trial .....	51
▪ Item selection.....	52
▪ Main survey coding .....	54
<b>CHAPTER 3 CONTEXT QUESTIONNAIRE DEVELOPMENT .....</b> 57	
<b>Introduction.....</b>	58
<b>The PISA context questionnaire framework .....</b>	58
<b>The PISA 2015 context questionnaires .....</b>	60
▪ The Student Questionnaire (computer-based and paper-based) .....	60

▪ The School Questionnaire (computer-based and paper-based) .....	60
▪ The Educational Career Questionnaire (computer-based).....	61
▪ The ICT Familiarity Questionnaire (computer-based).....	61
▪ The Parent Questionnaire (paper-based).....	61
▪ The Teacher Questionnaire .....	61
<b>Quality assurance in the development of questionnaires.....</b>	<b>62</b>
▪ National review.....	62
▪ Cognitive labs .....	62
▪ Translatability assessment .....	63
▪ Centralised trend material transfer.....	63
▪ Adaptation negotiation and verification .....	63
<b>CHAPTER 4 SAMPLE DESIGN .....</b>	<b>65</b>
<b>Target population and overview of the sampling design.....</b>	<b>66</b>
<b>Population coverage, and school and student participation rate standards .....</b>	<b>67</b>
▪ Coverage of the PISA international target population .....	67
▪ Accuracy and precision.....	68
▪ School response rates.....	68
▪ Student response rates .....	70
<b>Main study school sample .....</b>	<b>70</b>
▪ Definition of the national target population .....	70
▪ The sampling frame .....	70
▪ Stratification .....	71
▪ Assigning a measure of size to each school .....	74
▪ School sample selection .....	74
▪ Special school sampling situations .....	77
▪ PISA and national study overlap control .....	79
▪ Monitoring school sampling.....	80
<b>Student samples .....</b>	<b>84</b>
▪ Preparing a list of age-eligible students .....	85
▪ Selecting the student sample .....	85
▪ Preparing instructions for excluding students .....	85
▪ Sending the student tracking form to the school co-ordinator and test administrator .....	86
<b>Teacher samples .....</b>	<b>86</b>
<b>Definition of school .....</b>	<b>86</b>
<b>CHAPTER 5 TRANSLATION AND VERIFICATION OF THE SURVEY MATERIAL.....</b>	<b>91</b>
<b>Introduction .....</b>	<b>92</b>
<b>Development of source versions .....</b>	<b>92</b>
▪ Translatability assessment .....	92
▪ Production of the second source version in French.....	93
▪ Double translation from two source languages .....	93
<b>PISA translation and adaptation guidelines .....</b>	<b>94</b>
<b>Translation training sessions .....</b>	<b>94</b>
<b>Testing languages and translation/adaptation procedures .....</b>	<b>94</b>
<b>Centralised management of changes in trend .....</b>	<b>96</b>
<b>Mode effect study (see Chapter 2).....</b>	<b>96</b>
<b>International verification of the national versions .....</b>	<b>96</b>
<b>Verification of new computer-based test units .....</b>	<b>97</b>
▪ Verification of homolinguistic versions.....	97



▪ Verification of paper-based test units and booklet shell .....	98
▪ Verification of questionnaires .....	98
▪ Verification of coding guides .....	99
▪ Main survey verification.....	99
▪ Quantitative analyses of verification outcomes.....	99
<b>CHAPTER 6 FIELD OPERATIONS.....</b>	<b>101</b>
<b>Overview of roles and responsibilities.....</b>	<b>102</b>
▪ National Project Managers .....	102
▪ School Co-ordinators .....	103
▪ Test Administrators.....	103
<b>The selection of the school sample.....</b>	<b>104</b>
<b>Preparation of test booklets, questionnaires, and manuals.....</b>	<b>104</b>
<b>The selection of the student sample.....</b>	<b>105</b>
<b>Packaging and shipping materials .....</b>	<b>106</b>
▪ Field operations procedures specific to paper-based assessment countries .....	106
▪ Field operations procedures specific to computer-based assessment countries .....	106
<b>Test administration.....</b>	<b>107</b>
▪ Steps for setting up CBA test administration.....	107
▪ Steps for setting up PBA test administration .....	107
▪ Administering and monitoring the test.....	107
<b>Receipt of materials at the national centre after testing.....</b>	<b>108</b>
<b>Main survey review .....</b>	<b>108</b>
<b>CHAPTER 7 PISA QUALITY MONITORING.....</b>	<b>111</b>
<b>Introduction.....</b>	<b>112</b>
<b>Field trial and main survey review questionnaires .....</b>	<b>112</b>
<b>National centre consultations .....</b>	<b>112</b>
<b>PISA quality monitor (PQM) visits .....</b>	<b>113</b>
▪ Information collected in PQM visits during test administration .....	113
<b>Data adjudication.....</b>	<b>114</b>
<b>CHAPTER 8 SURVEY WEIGHTING AND THE CALCULATION OF SAMPLING VARIANCE.....</b>	<b>115</b>
<b>Survey weighting .....</b>	<b>116</b>
▪ The school base weight .....	117
▪ The school base weight trimming factor.....	118
▪ The within-school base weight.....	118
▪ The school non-response adjustment.....	119
▪ The grade non-response adjustment.....	119
▪ The within school non-response adjustment.....	122
▪ Trimming the student weights .....	122
▪ National option students .....	123
▪ International options .....	123
<b>Calculating sampling variance .....</b>	<b>123</b>
▪ The balanced repeated replication variance estimator .....	123
▪ Reflecting weighting adjustments .....	125
▪ Formation of variance strata .....	125
▪ Countries and economies where all students were selected for PISA.....	125

<b>CHAPTER 9 SCALING PISA DATA .....</b>	<b>127</b>
<b>Overview .....</b>	<b>128</b>
<b>Data yield and data quality .....</b>	<b>128</b>
▪ Targeted sample size, routing and data yield .....	128
▪ Classical test theory statistics: item analysis .....	131
▪ Response time analyses .....	134
▪ Position effects .....	140
<b>The IRT models for scaling .....</b>	<b>141</b>
▪ Moving from the Rasch model and partial credit model to the two-parameter logistic model and generalised partial credit model .....	141
▪ Measurement invariance (mode effect) model .....	144
<b>Latent regression model and population modelling .....</b>	<b>145</b>
<b>Analysis of data with plausible values .....</b>	<b>147</b>
<b>Application of IRT and population models to PISA .....</b>	<b>149</b>
▪ National and international item calibration .....	149
▪ Handling of item-by-country/language and item-by-mode interactions .....	150
▪ Mode effect study in the 2015 field trial: identifying items with mode effects .....	152
▪ Dimensionality and scaling of science trend and new items .....	163
▪ Scaling of reading and mathematics .....	165
▪ Dimensionality and scaling of collaborative problem solving .....	166
▪ Scaling of financial literacy .....	170
▪ Developing common scales for the purpose of trends .....	171
▪ Population modelling in PISA 2015 .....	179
▪ Generating plausible values .....	181
<b>CHAPTER 10 DATA MANAGEMENT PROCEDURES .....</b>	<b>187</b>
<b>Introduction .....</b>	<b>188</b>
<b>Data management at the international and national level .....</b>	<b>188</b>
▪ Data management at the international level .....	188
▪ Data management at the national level .....	189
<b>The data management process and quality control .....</b>	<b>190</b>
▪ Pre-processing .....	191
▪ Initial database load into SQL server and the cleaning and verification software .....	191
▪ Integration .....	192
<b>Harmonisation .....</b>	<b>194</b>
▪ Overview of the workflow .....	194
▪ Harmonisation, or harmonised variables .....	194
<b>Validation .....</b>	<b>195</b>
▪ Validation overview .....	195
▪ Treatment of inconsistent and extreme values in PISA 2015 main survey data .....	195
<b>Scoring and derivation .....</b>	<b>195</b>
▪ Scoring overview .....	195
▪ Derived variables overview .....	196
<b>Deliverables .....</b>	<b>196</b>
▪ Preparing files for public use and analysis .....	196
▪ Files prepared for national centre data reviews .....	197
▪ Records included in and excluded from the database .....	197
▪ Categorising missing data .....	198
▪ Data management and confidentiality, variable suppressions .....	198



<b>CHAPTER 11 SAMPLING OUTCOMES .....</b>	<b>203</b>
Population coverage .....	204
School and student response rates.....	205
Teacher response rates.....	214
Design effects and effective sample sizes .....	215
Variability of the design effect.....	217
▪ Design effects in PISA for performance variables.....	218
<b>CHAPTER 12 SCALING OUTCOMES.....</b>	<b>225</b>
Results of the IRT scaling and population modeling .....	226
▪ Unique item parameter estimation and national item deletion.....	226
▪ Generating student scale scores and reliability of the PISA scales .....	231
Transforming the plausible values to PISA scales.....	233
Linking error .....	237
International characteristics of the item pool .....	237
▪ Test targeting .....	237
▪ Domain inter-correlations.....	246
▪ Science scale and subscales.....	248
<b>CHAPTER 13 CODING DESIGN, CODING PROCESS, AND CODER RELIABILITY STUDIES.....</b>	<b>251</b>
Introduction.....	252
Coding procedures .....	252
Coding preparation .....	253
▪ Recruitment of national coder teams .....	253
▪ International coder training.....	254
▪ National coder training provided by the National Centres.....	254
Coding design.....	254
▪ Within-country and across-country coder reliability.....	256
Coder reliability studies .....	257
▪ Domain-level proportion agreement.....	257
▪ Item-level proportion agreement.....	259
▪ Coding category distributions .....	261
<b>CHAPTER 14 DATA ADJUDICATION .....</b>	<b>263</b>
Introduction.....	264
▪ PISA 2015 Technical Standards .....	264
▪ Implementing the standards – quality assurance.....	264
▪ Information available for adjudication.....	265
▪ Data adjudication process .....	266
▪ Data adjudication .....	267
General outcomes .....	267
▪ Overview of response rate issues .....	267
▪ Detailed country comments .....	268
▪ Albania .....	269
▪ Argentina .....	269
▪ Australia .....	270
▪ Canada.....	270
▪ Denmark .....	270
▪ Estonia .....	270

▪ Italy .....	270
▪ Kazakhstan .....	271
▪ Latvia .....	271
▪ Lebanon .....	271
▪ Lithuania .....	271
▪ Luxembourg .....	271
▪ Malaysia .....	271
▪ Montenegro .....	271
▪ Netherlands .....	272
▪ New Zealand .....	272
▪ Norway .....	272
▪ Spain .....	272
▪ Sweden .....	272
▪ Trinidad and Tobago .....	272
▪ United Kingdom .....	273
▪ United States .....	273
<b>CHAPTER 15 PROFICIENCY SCALE CONSTRUCTION.....</b>	<b>275</b>
<b>Introduction.....</b>	<b>276</b>
<b>Development of the described scales .....</b>	<b>277</b>
▪ Classification of items .....	277
▪ Defining the overall proficiency scale .....	277
▪ Identifying possible subscales .....	278
▪ Developing an item map.....	279
<b>Defining the proficiency levels .....</b>	<b>279</b>
<b>Reporting the results for pisa science .....</b>	<b>281</b>
▪ Building an item map for science .....	281
▪ Defining levels of scientific literacy.....	283
<b>CHAPTER 16 SCALING PROCEDURES AND CONSTRUCT VALIDATION OF CONTEXT QUESTIONNAIRE DATA.....</b>	<b>289</b>
<b>Introduction.....</b>	<b>290</b>
<b>Scaling methodology and construct validation .....</b>	<b>290</b>
▪ Scaling procedures .....	290
▪ Regular scales (PISA 2015) .....	291
▪ Trend scales (PISA 2006 - PISA 2015).....	292
▪ Interpreting results from IRT scaling .....	293
▪ Construct validation .....	295
▪ Internal consistency .....	295
▪ Cross-country comparability .....	295
▪ Derived variables based on IRT Scaling .....	300
<b>School Questionnaire derived variables.....</b>	<b>321</b>
▪ Simple questionnaire indices .....	321
▪ Derived variables based on IRT Scaling .....	323
<b>Educational Career Questionnaire.....</b>	<b>327</b>
▪ Simple questionnaire indices .....	327
<b>ICT Familiarity Questionnaire.....</b>	<b>328</b>
▪ Simple questionnaire indices .....	328
▪ Derived variables based on IRT Scaling .....	329
<b>Parent Questionnaire.....</b>	<b>331</b>
▪ Derived variables based on IRT Scaling .....	332



<b>Teacher Questionnaires</b>	335
▪ Simple questionnaire indices	336
▪ Derived variables based on IRT Scaling	337
<b>The PISA index of economic, social and cultural status (ESCS)</b>	339
▪ Computation of ESCS	339
▪ Trends in ESCS	341
<b>CHAPTER 17 QUESTIONNAIRE DESIGN AND COMPUTER-BASED QUESTIONNAIRE PLATFORM</b>	345
<b>Introduction</b>	346
<b>General questionnaire process</b>	346
<b>Step 1: Master questionnaires design</b>	347
▪ Field trial questionnaire design	349
▪ Main survey questionnaire design	350
<b>Step 2: Master questionnaires authoring</b>	350
▪ Questionnaire authoring tool	351
▪ Consistency check rule	360
<b>Step 3: Creation of national questionnaires</b>	362
<b>Step 4: National questionnaire adaptation and translation</b>	362
<b>Step 5: National questionnaires quality check</b>	363
<b>Step 6: Preparation of national questionnaires for delivery</b>	363
<b>Step 7: Data collection and quality monitoring</b>	365
<b>Step 8: Completion of data collection</b>	367
<b>Development process overview and technical infrastructure</b>	367
<b>Conclusion</b>	367
<b>CHAPTER 18 COMPUTER-BASED TEXTS</b>	369
<b>Introduction</b>	370
<b>Item rendering</b>	370
<b>Translation and online item review</b>	370
<b>School computer requirements</b>	371
<b>System diagnostic</b>	371
<b>Test delivery system</b>	371
<b>Data capture and scoring student responses</b>	373
<b>Open ended coding system</b>	373
<b>CHAPTER 19 INTERNATIONAL DATA PRODUCTS</b>	375
<b>Public use files</b>	376
▪ Variables excluded or suppressed for some or all countries	376
▪ File names and content	376
▪ Variables used in sampling, weighting and merging	377
▪ Missing code conventions	377
<b>Codebooks for the PISA 2015 public use data files</b>	377
<b>Data compendia tables</b>	378
<b>Data analysis and software tools</b>	378
▪ PISA Data Explorer (PDX)	378
<b>International Database Analyzer</b>	380
<b>Population and quality check of the PISA Data Explorer</b>	381

<b>ANNEXES</b>	383
<b>Annex A</b> Main survey item pool classification .....	384
<b>Annex B</b> Contrast coding used in conditioning .....	427
<b>Annex C</b> Standard errors of means, sample sizes, school variance estimates, and other sampling outcomes.....	428
<b>Annex D</b> Mapping of ISCED to years.....	435
<b>Annex E</b> National household possession items .....	436
<b>Annex F</b> Technical standards for PISA 2015 .....	438
<b>Annex G</b> Common and unique item parameters in each domain, by countries and languages .....	454
<b>Annex H</b> Scalar or metric invariant trend items in each domain .....	455
<b>Annex I</b> PISA contractors, staff and consultants .....	456
<b>BOXES</b>	
Box 1.1 Key features of PISA 2015 .....	25
Box 4.1 Illustration of probability proportional to size (PPS) sampling.....	76
<b>FIGURES</b>	
Figure 2.1 Comparison of construct coverage in the 2000-2012 PISA design by major and minor domains .....	31
Figure 2.2 Approach used to balance major/minor domains in 2015 and beyond.....	31
Figure 2.3 Domain coverage for PISA 2015.....	34
Figure 2.4 Field trial computer-based assessment design, with collaborative problem solving .....	36
Figure 2.5 Overview of the PISA 2015 main survey integrated design .....	36
Figure 2.6 Main survey paper-based assessment design .....	37
Figure 2.7 Main survey paper-based assessment design .....	38
Figure 2.8 Main Study Computer-Based Assessment Design .....	38
Figure 2.9 Main survey computer-based assessment design.....	40
Figure 2.10 Main study computer-based assessment combinations of science clusters.....	41
Figure 2.11 Lookup table for random number "S": Assignment of science cluster combinations .....	41
Figure 2.12 Main survey UH form design.....	42
Figure 2.13 Matrix of collaborative problem solving skills for PISA 2015 .....	44
Figure 2.14 Paging navigation used in PISA 2015.....	45
Figure 2.15 Item counts (field trial and main survey) by domain and delivery mode.....	52
Figure 2.16 Science item counts by framework category .....	53
Figure 2.17 Collaborative problem solving item counts by framework category.....	53
Figure 3.1 Modular structure of the PISA 2015 questionnaire design .....	59
Figure 3.2 Constructs identified as core content in the PISA 2015 Questionnaire Framework.....	59
Figure 3.3 Overview of the 19 policy issues (modules) and their relation to the questionnaires .....	60
Figure 4.1 School response rate standards .....	69
Figure 5.1 Translation procedures reported by national centres in the translation plan .....	95
Figure 5.2 Sample of a test adaptation spreadsheet (TAS) from the PISA 2015 field trial .....	97
Figure 5.3 Sample of a questionnaire adaptation spreadsheet (QAS) from the PISA 2015 field trial.....	98
Figure 6.1 Timing of paper-based assessment.....	107
Figure 9.1 Sample yield for the participating countries with CBA/CPS format.....	131
Figure 9.2 Sample yield for the participating countries with CBA or PBA format .....	131
Figure 9.3 Median response time by item – Collaborative problem solving.....	137
Figure 9.4 Median response time by PV1 proficiency level – Science trend items.....	137



Figure 9.5	Median response time by PV1 proficiency level – Science new items.....	138
Figure 9.6	Median response time vs. country median score (PV1) – All science items (2 clusters) .....	138
Figure 9.7	Variability of time used in science .....	139
Figure 9.8	Item response curve for an item where the international item parameter is not appropriate for one group (example from a different ILSA) .....	151
Figure 9.9	Comparison of slope parameter estimates across paper-based (horizontal axis) and computer-based (vertical axis) assessment modes for the PISA 2015 field trial data .....	154
Figure 9.10	Comparison of difficulty parameter estimates across paper-based (horizontal axis) and computer-based (vertical axis) assessment modes for the PISA 2015 field trial data .....	154
Figure 9.11	Split of country means by assessment mode for mathematics.....	158
Figure 9.12	Split of country means by gender for mathematics .....	159
Figure 9.13	Split of country means by random school split for mathematics .....	159
Figure 9.14	Split of country means by assessment mode for science .....	160
Figure 9.15	Split of country means by gender for science.....	160
Figure 9.16	Split of country means by random school split for science .....	161
Figure 9.17	Split of country means by assessment mode for reading .....	161
Figure 9.18	Split of country means by gender for reading.....	162
Figure 9.19	Split of country means by random school split for reading .....	162
Figure 9.20	Correlation plot among new science items averaged across countries (46 countries).....	165
Figure 9.21	Correlation plot among collaborative problem solving items averaged across countries before treating them as composite items (31 countries).....	168
Figure 9.22	Correlation plot among collaborative problem solving items averaged across countries after treating them as composite items (42 countries) .....	169
Figure 9.23	Percentage of variance from principal component analyses (6 example countries).....	170
<hr/>		
Figure 10.1	Overview of the data management process.....	189
Figure 10.2	Overview of the delivery and pre-processing phase.....	191
Figure 10.3	Initial load of the National Centre database into SQL server for processing.....	192
Figure 10.4	Integration process overview .....	193
Figure 10.5	Harmonisation process overview.....	194
Figure 10.6	PISA 2015 range restriction rules for inconsistent and extreme values for main survey data .....	199
Figure 10.7	PISA 2015 main survey country/variable suppression list .....	201
<hr/>		
Figure 12.1	Frequencies of international (invariant) and unique item parameters in maths (note that frequencies were counted using item-by-group pairs).....	228
Figure 12.2	Frequencies of international (invariant) and unique item parameters in reading (note that frequencies were counted using item-by-group pairs).....	228
Figure 12.3	Frequencies of international (invariant) and unique item in trend science (note that frequencies were counted using item-by-group pairs).....	229
Figure 12.4	Frequencies of international (invariant) and unique item in new science (note that frequencies were counted using item-by-group pairs).....	229
Figure 12.5	Frequencies of international (invariant) and unique item in CPS (note that frequencies were counted using item-by-group pairs).....	230
Figure 12.6	Frequencies of international (invariant) and unique item in financial literacy (note that frequencies were counted using item-by-group pairs).....	230
Figure 12.7	Item RP62 values and distribution of PV1 in maths .....	239
Figure 12.8	Item RP62 values and distribution of PV1 in reading .....	240
Figure 12.9	Item RP62 values and distribution of PV1 in science .....	240
Figure 12.10	Item RP62 values and distribution of PV1 in financial literacy .....	241
Figure 12.11	Item RP62 values and distribution of PV1 in collaborative problem solving.....	241
Figure 12.12	Percentage of respondents per country/economy at each level of proficiency for maths .....	242
Figure 12.13	Percentage of respondents per country/economy at each level of proficiency for reading .....	243
Figure 12.14	Percentage of respondents per country/economy at each level of proficiency for science .....	244
Figure 12.15	Percentage of respondents per country/economy at each level of proficiency for financial literacy .....	245

Figure 12.16 Percentage of respondents per country/economy at each level of proficiency for CPS.....	246
Figure 14.1 Attained school response rates.....	267
Figure 15.1 Simplified relationship between items and students on a proficiency scale .....	277
Figure 15.2 Calculating the RP values used to define PISA proficiency levels .....	281
Figure 15.3 A map for selected science items .....	282
Figure 15.4 Summary descriptions of the seven proficiency levels on the scientific literacy scale .....	284
Figure 15.5 Summary descriptions of the proficiency levels on the scientific literacy subscale <i>Explain phenomena scientifically</i> .....	284
Figure 15.6 Summary descriptions of the seven proficiency levels on the scientific literacy subscale <i>Evaluate and design scientific enquiry</i> .....	285
Figure 15.7 Summary descriptions of the seven proficiency levels on the scientific literacy subscale <i>Interpret data and evidence scientifically</i> .....	286
 Figure 16.1 Item characteristic curves for a four-category item under the generalised partial credit model (GPCM) .....	294
Figure 16.2 Illustration of an increase of the slope parameter, $\alpha$ , on category response curves for a four-category item under the generalised partial credit model (GPCM) .....	294
Figure 16.3 Example of an RMSD-plot: distribution of the RMSD statistic across groups .....	296
Figure 16.4 Example of an RMSD-distribution for a very well fitting item across all groups: All RMSD values are less than 0.1 .....	297
Figure 16.5 Computation of ESCS in PISA 2015 .....	340
 Figure 17.1 PISA 2015 questionnaire life cycle.....	346
Figure 17.2 Field trial computer-based design for Student (StQ) and Teacher Questionnaires (TCQ).....	349
Figure 17.3 Main survey computer-based design for Student (StQ) and Teacher Questionnaires (TCQ) .....	350
Figure 17.4 Questionnaire platform home page .....	351
Figure 17.5 QAT main view (with a specific question SC002 as an example) .....	352
Figure 17.6 Organisation of the main view of the QAT editor .....	352
Figure 17.7 The expended view information .....	353
Figure 17.8 Preview of a question with the QAT editor .....	354
Figure 17.9 Information template .....	355
Figure 17.10 Exclusive choice template (technical name <i>simpleMultipleChoiceRadioButton</i> ).....	355
Figure 17.11 Multiple choice template (technical name <i>simpleMultipleChoiceCheckbox</i> ).....	356
Figure 17.12 List of exclusive choice (table layout) template (technical name <i>complexMultipleChoiceRadioButton</i> ).....	356
Figure 17.13 List of multiple choice (table layout) template (technical name <i>complexMultipleChoiceCheckbox</i> ).....	357
Figure 17.14 List of text inputs (+ pie chart) template (technical name <i>simpleFieldsList</i> ) .....	357
Figure 17.15 Multiple list of text inputs (table layout) (technical name <i>complexFieldsList</i> ).....	358
Figure 17.16 Scale question type template (technical name <i>slider</i> ) .....	358
Figure 17.17 Free text input template (technical name <i>textfield</i> ) .....	359
Figure 17.18 Forced choice template (technical name <i>multipleItems</i> ) .....	359
Figure 17.19 Drop down (technical name <i>simpleDropDown</i> ).....	360
Figure 17.20 Drop down (table layout) template (technical name <i>complexDropDown</i> ) .....	360
Figure 17.21 Consistency check rule template .....	361
Figure 17.22 Consistency check message .....	361
Figure 17.23 Routing rule template .....	361
Figure 17.24 Questionnaire platform – administrative view.....	364
Figure 17.25 Distribution of the PISA 2015 servers .....	365
Figure 17.26 Logged events .....	366
 Figure 19.1 PISA database population and quality control .....	381

**TABLES**

Table 1.1	PISA 2015 participants.....	24
Table 4.1	Stratification variables used in PISA 2015 .....	72
Table 4.2	Schedule of school sampling activities .....	80
Table 4.3	Sampling frame unit.....	87
Table 8.1	Non-response classes.....	120
Table 9.1	Test mode, sample size per country and language.....	129
Table 9.2	Example output for examining response distributions.....	132
Table 9.3	Example table providing summary item statistics.....	133
Table 9.4	Flagging criteria for items in the item analyses .....	134
Table 9.5	Items excluded from the IRT scaling based on classical item analyses or technical problems .....	134
Table 9.6	Percentage of response time outliers in domains of PISA 2015 Main Survey.....	135
Table 9.7	Item cluster response time (in minutes) descriptive statistics.....	136
Table 9.8	Cluster level response time by PV1 proficiency level (min).....	136
Table 9.9	PISA 2009 and 2012 PBA proportion correct across clusters and across countries .....	140
Table 9.10	PISA 2015 CBA proportion correct across clusters and across countries .....	140
Table 9.11	PISA 2015 CBA median cluster timing averaged across countries (in minutes) .....	141
Table 9.12	PISA 2015 CBA omission rates across clusters and across countries .....	141
Table 9.13	Example for use of plausible values to partitioning the error.....	148
Table 9.14	Example for use of plausible values to partitioning the error – sample error, measurement error and standard error based on the 10 PVs.....	148
Table 9.15	Distribution of the test items across PISA cycles and assessment modes by domain used in PISA 2015 item calibration (main survey).....	150
Table 9.16	Correlations of item difficulty and item slope parameters between paper-based and computer-based trend items within and across domains.....	155
Table 9.17	Measurement invariance assessment using mode effect models for the PISA field trial data, analysed separately for the domains of financial literacy, maths, reading and science .....	157
Table 9.18	Model selection criteria for the unidimensional and the two-dimensional IRT models for trend and new science items.....	163
Table 9.19	Combination of collaborative problem solving items of Units 101 and 105 to achieve fair scoring in the PISA 2015 field trial .....	166
Table 9.20	Comparison of two-parameter logistic/generalised partial credit models and bifactor model for 164 CPS items .....	167
Table 9.21	List of composite items based on residual analyses .....	168
Table 9.22	Number of Rasch model items retained in the hybrid/IBCI model .....	174
Table 9.23	Changes in model fit summary.....	174
Table 9.24	Comparison of the Rasch/ partial credit model and the two-parameter logistic /generalised partial credit model for new items in the PISA 2015 field trial.....	174
Table 9.25	Distribution of 85 trend and 99 new items to the science scales and subscales .....	181
Table 11.1	PISA target populations and samples.....	206
Table 11.2	PISA target populations and samples, by adjudicated regions.....	208
Table 11.3	Response rates before school replacement .....	209
Table 11.4	Response rates before school replacement, by adjudicated regions .....	210
Table 11.5	Response rates after school replacement .....	211
Table 11.6	Response rates after school replacement, by adjudicated regions .....	212
Table 11.7	Response rates, students within schools after school replacement .....	213
Table 11.8	Response rates, students within schools after school replacement, by adjudicated regions .....	214
Table 11.9	Science teacher response rates .....	214
Table 11.10	Non-science teacher response rates .....	215
Table 11.11	Standard errors for the PISA 2015 main domain scales.....	219
Table 11.12	Design effects and effective sample sizes for scientific literacy .....	220

Table 11.13	Design effects and effective sample sizes for mathematical literacy .....	221
Table 11.14	Design effects and effective sample sizes for reading literacy .....	222
Table 11.15	Design effects and effective sample sizes for collaborative problem solving.....	223
Table 11.16	Design effects and effective sample sizes for financial literacy .....	224
<hr/>		
Table 12.1	Items that were excluded from the IRT analyses .....	226
Table 12.2	Percentage of common and unique item parameters in each domain for PISA 2015 .....	227
Table 12.3	Example of table for item parameter estimates provided to the countries.....	231
Table 12.4	Reliabilities of the PISA cognitive domains and Science subscales overall countries.....	231
Table 12.5	National reliabilities for main cognitive domains .....	232
Table 12.6	PISA 2015 transformation coefficients.....	235
Table 12.7	Average plausible values (PVs) and resampling-based standard errors (SE) by country/economy for the PISA domains of science, reading, mathematics, financial literacy, and collaborative problem solving (CPS).....	235
Table 12.8	Robust link error (based on absolute pairwise differences statistic $S_n$ ) for comparisons of performance between PISA 2015 and previous assessments.....	237
Table 12.9	Item and respondent classification for each score boundary in mathematics .....	238
Table 12.10	Item and respondent classification for each score boundary in reading .....	238
Table 12.11	Item and respondent classification for each score boundary in science .....	238
Table 12.12	Item and respondent classification for each score boundary in financial literacy .....	238
Table 12.13	Item and respondent classification for each score boundary in CPS .....	238
Table 12.14	Domain inter-correlations .....	247
Table 12.15	National-level domain inter-correlations based on 10 PVs .....	247
Table 12.16	Estimated correlations among domains and science knowledge subscales .....	249
Table 12.17	Estimated correlations among domains and science Competency subscales.....	249
Table 12.18	Estimated correlations among domains and science System subscales .....	249
<hr/>		
Table 13.1	Number of cognitive items by domain, item format and coding method .....	252
Table 13.2	Number of CBA coders by domain and coding design .....	255
Table 13.3	Multiple coding in CBA standard coding design.....	255
Table 13.4	Number of PBA coders by domain and coding design.....	256
Table 13.5	Multiple coding in PBA standard coding design .....	256
Table 13.6	Summary of within-country and across-country agreement (%) per domain for CBA participants .....	258
Table 13.7	Summary of within-country and across-country agreement (%) per domain for PBA participants .....	260
Table 13.8	Percentages of CBA and PBA participants with a different number of items for which proportion agreement is lower than 85%.....	260
Table 13.9	Summary of proportion agreement across the PISA participants .....	261
Table 13.10	Percentage of coders whose coding category distributions on more than 20% of coded items were significantly different from other coders, averaged across CBA and PBA participants .....	261
Table 13.11	Percentages of participant × item pairs that have more than two coders' coding category distributions significantly different from other coders .....	261
<hr/>		
Table 15.1	Scientific literacy performance band definitions on the PISA scale .....	283
<hr/>		
Table 16.1	OECD mean and standard deviation (S.D.) for the untransformed WLEs of regular scales in the different PISA 2015 context questionnaires .....	291
Table 16.2	Scaling constants ( $A$ , $B$ ) and correlations between original and newly derived 2006 WLEs for trend scales in 2015 .....	293
Table 16.3	Derived variables in the PISA 2015 Student Questionnaire .....	297
Table 16.4	Indicators of household possessions and home background indices .....	300
Table 16.5	Scale reliabilities for Household possessions indices in OECD countries.....	301
Table 16.6	Scale reliabilities for Household possessions indices in partner countries and economies .....	302
Table 16.7	Item parameters for national home possession indicators in OECD countries .....	303
Table 16.8	Item parameters for national home possession indicators in partner countries and economies .....	304
Table 16.9	Item parameters for Home possessions (HOMEPOS) .....	304
Table 16.10	Item parameters for Family wealth (WEALTH) .....	305
Table 16.11	Item parameters for Cultural possessions at home (CULTPOSS) .....	305



Table 16.12	Item parameters for Home educational resources (HEDRES) .....	305
Table 16.13	Item parameters for ICT Resources (ICTRES).....	305
Table 16.14	Scale reliabilities for BELONG in OECD countries.....	306
Table 16.15	Scale reliabilities for BELONG in partner countries and economies .....	306
Table 16.16	Item parameters for Sense of Belonging to School (BELONG).....	307
Table 16.17	Scale reliabilities for COOPERATE and CPSVALUE in OECD countries .....	307
Table 16.18	Scale reliabilities for COOPERATE and CPSVALUE in partner countries and economies .....	308
Table 16.19	Item parameters for Enjoy co-operation (COOPERATE) .....	308
Table 16.20	Item parameters for Value co-operation (CPSVALUE).....	308
Table 16.21	Scale reliabilities for ENVAWARE and ENVOPT in OECD countries.....	309
Table 16.22	Scale reliabilities for ENVAWARE and ENVOPT in partner countries and economies .....	310
Table 16.23	Item parameters for Environmental Awareness (ENVAWARE).....	310
Table 16.24	Item parameters for Environmental optimism (ENVOPT) .....	310
Table 16.25	Scale reliabilities for JOYSCIE and INTBRSCI in OECD countries .....	311
Table 16.26	Scale reliabilities for JOYSCIE and INTBRSCI in partner countries and economies.....	311
Table 16.27	Item parameters for Enjoyment of science (JOYSCIE) .....	312
Table 16.28	Item parameters for Interest in broad science topics (INTBRSCI).....	312
Table 16.29	Scale reliabilities for all seven indices relating to Science learning in school in OECD countries .....	313
Table 16.30	Scale reliabilities for all seven indices relating to Science learning in school in partner countries and economies .....	313
Table 16.31	Item parameters for Disciplinary climate in science classes (DISCLISCI).....	314
Table 16.32	Item parameters for Inquiry-based science teaching and learning practices (IBTEACH) .....	314
Table 16.33	Item parameters for Teacher support in a science classes (TEACHSUP).....	314
Table 16.34	Item parameters for Teacher-directed science instruction (TDTEACH).....	314
Table 16.35	Item parameters for Perceived Feedback (PERFEED) .....	315
Table 16.36	Item parameters for Adaption of instruction (ADINST).....	315
Table 16.37	Item parameters for Instrumental motivation (INSTSCIE).....	315
Table 16.38	Scale reliabilities for ANXTEST and MOTIVAT in OECD countries .....	316
Table 16.39	Scale reliabilities for ANXTEST and MOTIVAT in partner countries and economies.....	316
Table 16.40	Item parameters for Test Anxiety (ANXTEST).....	317
Table 16.41	Item parameters for Achievement motivation (MOTIVAT) .....	317
Table 16.42	Scale reliabilities for the Parental support index in OECD countries .....	317
Table 16.43	Scale reliabilities for the Parental support index in partner countries and economies .....	318
Table 16.44	Item parameters for Parents emotional support (EMOSUPS).....	318
Table 16.45	Scale reliabilities for indices on Science related dispositions in OECD countries.....	319
Table 16.46	Scale reliabilities for indices on Science related dispositions in partner countries and economies .....	319
Table 16.47	Item parameters for Science self-efficacy (SCIEEFF) .....	320
Table 16.48	Item parameters for Epistemological beliefs (EPIST) .....	320
Table 16.49	Item parameters for Science activities (SCIEACT).....	320
Table 16.50	Derived variables in the PISA 2015 School Questionnaire .....	321
Table 16.51	Scale reliabilities for School Questionnaire indices in OECD countries.....	323
Table 16.52	Scale reliabilities for School Questionnaire in partner countries and economies .....	324
Table 16.53	Item parameters for Educational leadership (LEAD) .....	325
Table 16.54	Item parameters for Curricular development (LEADCOM).....	325
Table 16.55	Item parameters for Instructional leadership (LEADINST) .....	325
Table 16.56	Item parameters for Professional development (LEADPD).....	326
Table 16.57	Item parameters for Teachers participation (LEADTCH) .....	326
Table 16.58	Item parameters for Shortage of educational material (EDUSHORT).....	326
Table 16.59	Item parameters for Shortage of educational staff (STAFFSHORT).....	326
Table 16.60	Item parameters for Student-related factors affecting school climate (STUBEHA).....	327
Table 16.61	Item parameters for Teacher-related factors affecting school climate (TEACHBEHA) .....	327
Table 16.62	Derived variables in the optional PISA 2015 Educational Career Questionnaire.....	327
Table 16.63	Derived variables in the optional PISA 2015 ICT Familiarity Questionnaire.....	328
Table 16.64	Scale reliabilities for ICT Familiarity Questionnaire indices in OECD countries.....	329

Table 16.65	Scale reliabilities for ICT Familiarity Questionnaire in partner countries and economies .....	329
Table 16.66	Item parameters for ICT use outside of school for leisure (ENTUSE).....	330
Table 16.67	Item parameters for ICT use outside of school for schoolwork (HOMESCH) .....	330
Table 16.68	Item parameters for Use of ICT at school in general (USESCH).....	330
Table 16.69	Item parameters for Students' ICT Interest (INTICT) .....	331
Table 16.70	Item parameters for Students' Perceived ICT Competence (COMPICT) .....	331
Table 16.71	Item parameters for Students' Perceived Autonomy related to ICT Use (AUTICT).....	331
Table 16.72	Item parameters for Students' ICT as a topic in Social Interaction (SOIAICT) .....	331
Table 16.73	Derived variables in the optional PISA 2015 Parent Questionnaire.....	332
Table 16.74	Scale reliabilities for the Parent Questionnaire indices in OECD countries .....	332
Table 16.75	Scale reliabilities for the Parent Questionnaire in partner countries and economies .....	332
Table 16.76	Item parameters for Child's past science activities (PRESUPP).....	333
Table 16.77	Item parameters for Parental current support for learning at home (CURSUPP).....	333
Table 16.78	Item parameters for Parental emotional support (EMOSUPP).....	333
Table 16.79	Item parameters for School policies for parental involvement (PASCHPOL).....	334
Table 16.80	Item parameters for Parents perceived school quality (PQSCHOOL).....	334
Table 16.81	Item parameters for Parents' view on science (PQGENSCI).....	334
Table 16.82	Item parameters for Parents concerns regarding environmental topics (PQENPERC) .....	335
Table 16.83	Item parameters for Parents' view on future environmental topics (PQENVOPT) .....	335
Table 16.84	Derived variables in the optional PISA 2015 Teacher Questionnaire .....	335
Table 16.85	Scale reliabilities for Teacher Questionnaire indices in OECD countries.....	337
Table 16.86	Scale reliabilities for Teacher Questionnaire indices in partner countries and economies .....	337
Table 16.87	Item parameters for Satisfaction with the current job environment (SATJOB) .....	337
Table 16.88	Item parameters for Satisfaction with teaching profession (SATTEACH) .....	337
Table 16.89	Item parameters for Transformational leadership teachers view (TCLEAD) .....	338
Table 16.90	Item parameters for Educational material shortage teachers view (TCEDUSHORT).....	338
Table 16.91	Item parameters for Staff shortage teachers view (TCSTAFFSHORT).....	338
Table 16.92	Item parameters for Science teacher collaboration (COLSCIT).....	338
Table 16.93	Item parameters for Exchange and co-ordination for teaching (EXCHT) .....	339
Table 16.94	Item parameters for Self-efficacy related to teaching science content (SETEACH) .....	339
Table 16.95	Item parameters for Self-efficacy related to science content (SECONT).....	339
Table 16.96	Factor loadings and reliability (Cronbach's Alpha) of ESCS 2015 in OECD countries .....	340
Table 16.97	Factor loadings and reliability (Cronbach's Alpha) of ESCS 2015 in partner countries and economies .....	341
Table 17.1	PISA 2015 questionnaires .....	346
Table 17.2	Questionnaire participation in PISA 2015 main survey .....	347
Table 19.1	Robust link error for comparisons of performance between PISA 2015 and previous assessments.....	380
Table A.1	PISA 2015 main survey trend science item classification.....	384
Table A.2	PISA 2015 main survey new science item classification .....	394
Table A.3	PISA 2015 main survey trend reading item classification.....	402
Table A.4	PISA 2015 main survey trend math item classification .....	410
Table A.5	PISA 2015 main survey financial literacy item classification .....	418
Table A.6	PISA 2015 main survey CPS item classification .....	421
Table C.1	Standard errors of the student performance mean estimate by country and by domain.....	428
Table C.2	Sample sizes by country and by domain .....	429
Table C.3	School variance estimate by country and by domain.....	431
Table C.4	Intraclass correlation by country and by domain .....	432
Table C.5	Within explicit strata intraclass correlation by country and by domain .....	433
Table C.6	Percentage of school variance explained by explicit stratification variables by country and by domain .....	434
Table D.1	Mapping of ISCED to years .....	435
Table E.1	National household possession items.....	436



# Reader's Guide

## Country coverage

This publication features data on 72 countries and economies, including all 35 OECD countries and 37 partner countries and economies.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

B-S-J-G (China) refers to the four PISA-participating Chinese provinces: Beijing, Shanghai, Jiangsu and Guangdong.

CABA (Argentina) refers to the Ciudad Autónoma de Buenos Aires, Argentina.

FYROM refers to the Former Yugoslav Republic of Macedonia.

Russia refers to the Russian Federation.

## List of abbreviations – the following abbreviations are used in this report:

A2PLM:	Two-Parameter Logistic Model	IALS:	International Adult Literacy Survey
ACER:	Australian Council for Educational Research	IBCI:	Item-by-Country Interactions
AIC:	Akaike information criterion	ICC:	Item Characteristic Curve
aSPe:	University of Liege, Belgium	ICF:	Item Characteristic Function
BAS:	Booklet Adaptation Spreadsheet	ICR:	Inter-Country Coder Reliability Study
BIC:	Bayesian information criterion	ICT:	Information Communication Technology
BRR:	Balanced Repeated Replication	IEA:	International Association for the Evaluation of Educational Achievement
CBA:	Computer Based Assessment	ILS:	University of Oslo, Norway
CITO:	National Institute for Educational Measurement, the Netherlands	ILSA:	International Large Scale Assessment
CPS:	Collaborative Problem Solving	INES:	OECD Indicators of Education Systems
DIF:	Differential Item Functioning	INT:	International
DIPF:	The German Institute for International Educational Research	IPN:	Leibniz Institute for Science and Mathematics Education, Germany
DTCS:	DRA Target Cluster Size	IRT:	Item Response Theory
ENR:	Enrolment of 15-year-olds	ISCED:	International Standard Classification of Education
ESCS:	PISA Index of Educational, Social and Cultural Status	ISCO:	International Standard Classification of Occupations
ETCS:	CBA Tagert Cluster Size	ISEI:	International Socio-Economic Index
ETS:	Educational Testing Service	MAS:	Manuals Adaptation Spreadsheets
FL:	Financial Literacy	MEG:	Mathematics Expert Group
FT:	Field Trial	MENR:	Enrolment for moderately small school
FOC:	Final Optical Check	MCMLM:	Mixed-coefficients multinomial logit model
GPCM:	Generalised Partial Credit Model	MD:	Mean Deviation
I:	Sampling Interval		

MNSQ:	Mean Square	RN:	Random Number
MOS:	Measure of Size	R POLY:	R-Polyserial
MS:	Main Survey	RP:	Response Probability
NCQM:	National Centre Quality Monitor	SC:	School Co-ordinator
NEP:	National Enrolled Population	S.E.:	Standard Error
NIER:	National Institute for Educational Research, Japan	SEN:	Special Education Needs
NPM:	National Project Manager	S.D.:	Standard Deviation
OLT:	Open Language Tool	SJT:	Situational Judgment Tests
PBA:	Paper Based Assessment	SPT:	Study Programme Table
PCA:	Principal Component Analysis	TA:	Test Administrator
PPS:	Probability Proportional to Size	TAG:	Technical Advisory Group
PGB:	PISA Governing Board	TAS:	Test Adaptation Spreadsheet
PCM:	Partical Credit Model	TCS:	Target Cluster Size
PQM:	PISA Quality Monitor	TIMSS:	Third International Mathematics and Science Study
PV:	Plausible Values	TMS:	Translation Management System
QAS:	Questionnaire Adaptations Spreadsheet	UH:	Une Heure booklet
R BIS:	R-Biserial	VENR:	Enrolment for very small schools
RMSD:	Root Mean Square Deviation	WLE:	Weighted Likelihood Estimates

**List of country codes –** the following country codes are used in some tables in this report:

OECD countries	ISO code	OECD countries	ISO code
Australia	AUS	Korea	KOR
Austria	AUT	Latvia	LVA
Belgium	BEL	Luxembourg	LUX
Canada	CAN	Mexico	MEX
Chile	CHL	Netherlands	NLD
Czech Republic	CZE	New Zealand	NZL
Denmark	DNK	Norway	NOR
Estonia	EST	Poland	POL
Finland	FIN	Portugal	PRT
France	FRA	Slovak Republic	SVK
Germany	DEU	Slovenia	SVN
Greece	GRC	Spain	ESP
Hungary	HUN	Sweden	SWE
Iceland	ISL	Switzerland	CHE
Ireland	IRL	Turkey	TUR
Israel	ISR	United Kingdom	GBR
Italy	ITA	United States	USA
Japan	JPN		



<b>Partner countries and economies</b>	<b>ISO code</b>	<b>Partner countries and economies</b>	<b>ISO code</b>
Albania	ALB	Lithuania	LTU
Algeria	DZA	Macao (China)	MAC
Argentina	ARG	Malaysia	MYS
Brazil	BRA	Malta	MLT
Bulgaria	BGR	Moldova	MDA
B-S-J-G (China)	QCH	Montenegro	MNE
Colombia	COL	Peru	PER
Costa Rica	CRI	Qatar	QAT
Croatia	HRV	Romania	ROU
Cyprus <sup>1</sup>	CYP	Russia	RUS
Dominican Republic	DOM	Singapore	SGP
FYROM	MKD	Chinese Taipei	TAP
Georgia	GEO	Thailand	THA
Hong Kong (China)	HKG	Trinidad and Tobago	TTO
Indonesia	IDN	Tunisia	TUN
Jordan	JOR	United Arab Emirates	ARE
Kazakhstan	KAZ	Uruguay	URY
Kosovo	KSV	Viet Nam	VNM
Lebanon	LBN		

1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

The reader should note that a series of technical documents are available from the PISA website: [www.oecd.org/pisa](http://www.oecd.org/pisa).





1

# Programme for International Student Assessment: an overview

<b>Introduction .....</b>	22
<b>Participation .....</b>	23
<b>Features of PISA.....</b>	23
<b>Technical innovations in PISA 2015.....</b>	25
<b>Managing and implementing PISA .....</b>	26
<b>PISA 2015 publications .....</b>	27

## INTRODUCTION

The OECD Programme for International Student Assessment (PISA) is a collaborative effort among OECD member countries to measure how well 15-year-old students approaching the end of compulsory schooling are prepared to meet the challenges of today's knowledge societies. The assessment is forward-looking: rather than focusing on the extent to which these students have mastered a specific school curriculum, it looks at their ability to use their knowledge and skills to meet real-life challenges. This orientation reflects a change in curricular goals and objectives, focusing more on what students can do with what they learn at school.

PISA surveys take place every three years. The first survey took place in 2000 (followed by a further 8 and 3 countries and economies in 2001 and 2002, respectively), the second in 2003, the third in 2006, the fourth in 2009 (followed by a further 10 countries and economies in 2010), the fifth in 2012 and the sixth in 2015. The results of these surveys have been published in a series of reports (OECD, 2017a-b-c, 2016a-b, 2014a-b-c, 2013a-b-c, 2011, 2010a-b-c-d-e, 2007, 2004, 2001; OECD/UNESCO Institute for Statistics (2003); and Walker (2011)) and a wide range of thematic and technical reports. The next survey will occur in 2018. For each assessment, reading, mathematics or science is chosen as the major domain and given greater emphasis than the remaining two minor domains. In 2000 and 2009 the major domain was reading; in 2003 and 2012 it was mathematics, and in 2006 and 2015 it was science.

PISA is an age-based survey, assessing 15-year-old students in school in grade 7 or higher. These students are approaching the end of compulsory schooling in most participating countries, and school enrolment at this level is close to universal in almost all OECD countries.

The PISA assessments take a literacy perspective, focusing on the extent to which students can apply the knowledge and skills they have learned and practised at school when confronted with situations and challenges for which that knowledge may be relevant. That is, PISA assesses the extent to which students can use their reading skills to understand and interpret the various kinds of written material that they are likely to meet as they navigate everyday life; the extent to which students can use their mathematical knowledge and skills to solve various kinds of numerical and spatial challenges and problems; and the extent to which students can use their scientific knowledge and skills to understand, interpret and resolve various kinds of scientific situations and challenges. The PISA 2015 domains are fully defined in *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving* (OECD, 2017d).

PISA also conducts assessments of additional cross-curricular competencies from time to time as participating countries see fit. For example, in PISA 2003, an assessment of general problem-solving competencies was included and in PISA 2009 a computer-delivered digital reading assessment (DRA) was included for the first time. In PISA 2012 a computer-delivered assessment of mathematics and problem solving was added, along with an assessment of financial literacy. The DRA was included again in 2012. In PISA 2015 financial literacy was assessed for a second time but for this cycle in computer-based form. A computer-based assessment of collaborative problem solving was also added.

In addition, PISA uses Student Questionnaires to collect information from students on various aspects of their home, family and school background, and School Questionnaires to collect information from schools about various aspects of organisation and educational provision in schools. There are also optional questionnaire modules for students asking about Familiarity with Information and Communications Technology (ICT) about aspects of their Educational Career (EC). In PISA 2015, 18 countries also administered a Parent Questionnaire to the parents of the students participating in PISA. A Teacher Questionnaire was also developed for the first time in PISA and this was administered in 19 countries. Chapter 17 provides information about participation in the optional questionnaires.

Using the data from questionnaires, analyses linking contextual information with student achievement can address:

- differences between countries in the relationships between student-level factors (such as gender and socio-economic background) and achievement
- differences in the relationships between school-level factors and achievement across countries
- differences in the proportion of variation in achievement between (rather than within) schools, and differences in this value across countries
- differences between countries in the extent to which schools moderate or increase the effects of individual-level student factors and student achievement



- differences in education systems and national context that are related to differences in student achievement across countries
- changes in any or all of these relationships over time by linking PISA 2000, PISA 2003, PISA 2006, PISA 2009 and PISA 2012.

By collecting such information at the student and school level on a cross-nationally comparable basis, PISA adds significantly to the knowledge base that is available from national official statistics, such as aggregate national statistics on the educational programmes completed and the qualifications obtained by individuals.

The framework for the PISA 2015 questionnaires is included in *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving* (OECD, 2017d).

## PARTICIPATION

The first PISA survey was conducted in 2000 in 32 countries and economies (including 29 OECD member countries) using written tasks answered in schools under independently supervised test conditions. Another 11 countries and economies completed the same assessment in 2001 and 2002. PISA 2000 surveyed reading, mathematics and science, with a primary focus on reading.

The second PISA survey, conducted in 2003 in 41 countries and economies, assessed reading, mathematics and science, and problem solving with a primary focus on mathematics.

The third survey covered reading, mathematics and science, with a primary focus on science, and was conducted in 2006 in 57 countries and economies.

PISA 2009, the fourth PISA survey covered reading, mathematics and science, with a primary focus on reading, and was conducted in 65 countries and economies. Another 10 additional participants completed the PISA 2009 assessment in 2010.

PISA 2012, the fifth PISA survey covered reading, mathematics, science, problem solving and financial literacy with a primary focus on mathematics, and was conducted in 35 OECD countries and 30 partner countries and economies.

PISA 2015, the sixth PISA survey covered reading, mathematics, science, collaborative problem solving and financial literacy with a primary focus on science, and was conducted in 35 OECD countries and 37 partner countries and economies.

The participants in PISA 2015 are listed in Table 1.1. The table also indicates whether countries/economies participated in the computer-based mode (CBA) or paper-based mode (PBA), and shows the countries and economies that participated in the collaborative problem solving (CPS) and/or financial literacy assessments.

## FEATURES OF PISA

The technical characteristics of the PISA survey involve a number of different aspects:

- the design of the tests and questionnaires and the features incorporated in the instruments developed for PISA
- the sampling design, including both the school sampling and the student sampling requirements and procedures
- rules and procedures to guarantee the equivalence of the different language versions used within and between participating countries and economies, and taking into account the diverse cultural contexts of those countries and economies
- various operational procedures, including test administration arrangements, data capture and processing, and quality assurance mechanisms designed to ensure the generation of comparable data from all countries and economies
- the technical requirements and procedures for administering computer-based tests in schools
- scaling and analysis of the data and their subsequent reporting
- quality assurance procedures that enable PISA to provide high quality data to support policy formation and review.

This report describes the above-mentioned methodologies as they have been implemented in PISA 2015. Box 1.1 provides an overview of the central design elements of PISA 2015.

Table 1.1 PISA 2015 participants

	Mode	CPS	Financial literacy
<b>OECD countries</b>			
Australia	CBA	Yes	Yes
Austria	CBA	Yes	No
Belgium	CBA	Yes	Yes (Flemish community only)
Canada	CBA	Yes	Yes (7 provinces)
Chile	CBA	Yes	Yes
Czech Republic	CBA	Yes	No
Denmark	CBA	Yes	No
Estonia	CBA	Yes	No
Finland	CBA	Yes	No
France	CBA	Yes	No
Germany	CBA	Yes	No
Greece	CBA	Yes	No
Hungary	CBA	Yes	No
Iceland	CBA	Yes	No
Ireland	CBA	No	No
Israel	CBA	Yes	No
Italy	CBA	Yes	Yes
Japan	CBA	Yes	No
Korea	CBA	Yes	No
Latvia	CBA	Yes	No
Luxembourg	CBA	Yes	No
Mexico	CBA	Yes	No
Netherlands	CBA	Yes	Yes
New Zealand	CBA	Yes	No
Norway	CBA	Yes	No
Poland	CBA	No	Yes
Portugal	CBA	Yes	No
Slovak Republic	CBA	Yes	Yes
Slovenia	CBA	Yes	No
Spain	CBA	Yes	Yes
Sweden	CBA	Yes	No
Switzerland	CBA	No	No
Turkey	CBA	Yes	No
United Kingdom	CBA	Yes	No
United States	CBA	Yes	Yes
<b>Partner countries/economies</b>			
Albania	PBA	No	No
Algeria	PBA	No	No
Argentina	PBA	No	No
Brazil	CBA	Yes	Yes
B-S-J-G (China) <sup>1</sup>	CBA	Yes	Yes
Bulgaria	CBA	Yes	No
Colombia	CBA	Yes	No
Costa Rica	CBA	Yes	No
Croatia	CBA	Yes	No
Cyprus <sup>2</sup>	CBA	Yes	No
Dominican Republic	CBA	No	No
FYROM <sup>3</sup>	PBA	No	No
Georgia	PBA	No	No
Hong Kong (China)	CBA	Yes	No
Indonesia	PBA	No	No
Jordan	PBA	No	No
Kazakhstan	PBA	No	No
Kosovo	PBA	No	No
Lebanon	PBA	No	No
Lithuania	CBA	Yes	Yes
Macao (China)	CBA	Yes	No
Malaysia	CBA	Yes	No
Malta	PBA	No	No
Moldova	PBA	No	No
Montenegro	CBA	Yes	No
Peru	CBA	Yes	Yes
Qatar	CBA	No	No
Romania	PBA	No	No
Russia	CBA	Yes	Yes
Singapore	CBA	Yes	No
Chinese Taipei	CBA	Yes	No
Thailand	CBA	Yes	No
Trinidad and Tobago	PBA	No	No
Tunisia	CBA	Yes	No
United Arab Emirates	CBA	Yes	No
Uruguay	CBA	Yes	No
Viet Nam	PBA	No	No

1. B-S-J-G (China) refers to the four PISA-participating China provinces: Beijing, Shanghai, Jiangsu and Guangdong.

2. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

3. FYROM refers to the Former Republic of Yugoslavia.



### Box 1.1 Key features of PISA 2015

#### **The content**

The PISA 2015 survey focused on science, with reading and mathematics as minor areas of assessment. PISA 2015 also included the assessment of an innovative domain, collaborative problem solving and the assessment of financial literacy which was optional for countries and economies.

PISA assesses not only whether students can reproduce knowledge, but also whether they can extrapolate from what they have learned and apply their knowledge in new situations. It emphasises the mastery of processes, the understanding of concepts, and the ability to function in various types of situations.

#### **The students**

Approximately 540 000 students completed the assessment in 2015, representing about 29 million 15-year-olds in the schools of the 72 participating countries and economies.

#### **The assessment**

For the first time in PISA 2015, computer-based tests were the main mode of assessment. Paper-based alternatives were used in 15 countries and economies that did not have the resources available for computer-based testing in schools. The tests lasted a total of two hours for each student and covered reading, science, mathematics and collaborative problem solving in the majority of computer-based countries and economies<sup>1</sup> and reading, science and mathematics in paper-based countries and economies. An additional 60 minutes were devoted to the computer-based assessment of financial literacy in countries and economies that chose to implement this option.

Test items were a mixture of multiple-choice items and questions requiring students to construct their own responses. The items were organised in groups based on a text or graphic setting out a real-life situation. Some science tasks presented students with an interactive scenario (e.g. a science experiment) which required manipulation of elements within the scenario, while collaborative problem solving was assessed via interactive chat-based tasks with branching based on student responses. A total of 810 minutes of test items in reading, science, mathematics and collaborative problem solving were included, with different students taking different combinations of test items.

Students answered a background questionnaire, which took around 30 minutes to complete. The questionnaire sought information about themselves, their homes and their school and learning experiences. School principals completed a questionnaire that covered the school system and the learning environment. In some countries and economies, optional questionnaires were distributed to parents, who were asked to provide information on their perceptions of and involvement in their child's school, their support for learning in the home, and their child's career expectations, particularly in science-based occupations. Countries and economies could choose two other optional questionnaires for students: one asked students about their familiarity with and use of information and communication technologies, and the second sought information about their education to date. For the first time in PISA 2015 countries and economies could also opt to distribute a questionnaire to teachers.

1. The test of collaborative problem solving was not available in paper-based format and a small number of computer-based countries (the Dominican Republic, Ireland, Poland, Qatar and Switzerland) also chose not to administer this part of the assessment.

## **TECHNICAL INNOVATIONS IN PISA 2015**

A major innovation in PISA 2015 was the move from a primarily paper-based survey that included optional computer-based modules to a fully computer-delivered survey. A paper-based version of the assessment that included only trend units was developed for the small number of countries and economies that did not implement the computer-based survey (see Figure 1.1). The computer-based delivery mode made it possible to measure new and expanded aspects of the domain constructs. In particular, the addition of interactive tasks in science allowed students to manipulate variables in simulated scientific enquiries, and the collaborative problem solving assessment applied interactive chat-based tasks with branching based on student responses. Chapter 2 describes these tasks in more detail and Chapter 18 describes the technical aspects of the computer delivery platform. Chapter 17 describes the platform used for the development and delivery of background questionnaires for students, school principals and teachers.

In addition to the development of computer-based delivery in schools, an interactive portal was set up to support survey implementation and enhance communication between national teams and the international contractors. Chapter 6

describes the use of this portal for a variety of tasks while Chapter 18 describes the technical aspects of the portal. Chapter 5 describes the use of the online portal for translation and adaptation procedures in more detail.

A further development of computer-based activities was onscreen marking of tests which was an option for national centres in previous PISA cycles but became the main medium for test marking in PISA 2018. This offered considerable advantages in monitoring marking activities and enabling real-time checks on marker reliability, thereby increasing the accuracy and reliability of marking open-ended responses. In addition, responses from closed items in test and questionnaires were captured automatically without the need for data entry, saving time and avoiding potential operator error. Chapter 13 describes the marking process while Chapter 18 describes technical details of the Open-Ended Coding System (OECS) and the direct capture of responses from closed items.

The move to computer-based delivery as the main mode of assessment also made it possible to collect more in depth information not just on student responses but also the process behind those responses, such as the amount of time it took to complete each task and the number of actions taken by the student. Chapter 18 describes the type of information which was collected.

There were also innovations in the scaling model used and in the measurement of trends across PISA cycles. The ability to establish and maintain trends over time is an important goal for PISA. In PISA 2015 the assessment design was enhanced to increase coverage of minor domains, with the aim of strengthening trend measurement. The integrated design for the assessment which is described in Chapter 2 increased the number of items for the minor domains to previous major domain levels, reducing the potential for introducing systematic measurement error across PISA cycles. The methodology incorporated all available data from previous cycles for scaling and analysis, thus providing a solid base for linking across cycles and between paper-based and computer-based administrations.

PISA, as with other large scale international studies, uses an Item Response Theory (IRT) approach in the analysis and scaling of the data and the measurement of trends across cycles. The IRT model used in PISA 2015 underwent some modifications compared with previous cycles which based the scaling entirely on a Rasch model. To increase the ability of the scaling to address the complexities of PISA data, PISA 2015 implemented a hybrid model which combined a Rasch approach with other IRT models, with a two-parameter-logistic model and a generalised partial credit model (GPCM) used where appropriate. Chapter 9 describes this innovative approach in detail and Chapter 12 presents scaling outcomes.

## **MANAGING AND IMPLEMENTING PISA**

PISA is implemented within a framework established by the PISA Governing Board (PGB) which includes representation from all participating countries and economies at senior policy levels. The PGB establishes policy priorities and standards for developing indicators, for establishing assessment instruments, and for reporting results. Annex G lists the members of the PISA Governing Board and the observers from partner countries and economies.

Experts from participating countries and economies served on working groups linking the programme policy objectives with the best internationally available technical expertise in the assessment areas and in the areas which were included in the context questionnaires. These expert groups were referred to as Subject Matter Expert Groups (EGs) and the Questionnaire Expert Group (QEG). By participating in these expert groups and regularly reviewing outcomes of the groups' meetings, countries and economies ensured that the instruments were internationally valid, that they took the cultural and educational contexts of participating countries and economies into account, that the assessment materials had strong measurement potential, and that the instruments emphasised authenticity and educational validity. See Annex G for the list of members of the expert groups.

Each of the participating countries and economies appointed a National Project Manager (NPM) to implement PISA nationally. The NPMs ensured that internationally agreed common technical and administrative procedures were employed. These managers played a vital role in developing and validating the international assessment instruments and ensured that PISA implementation was of high quality. The NPMs also contributed to the verification and evaluation of the survey results, analyses and reports. Annex G also lists the PISA 2015 NPMs.

The OECD Secretariat was responsible for the overall management of the programme. It monitored its implementation on a day-to-day basis, served as the secretariat for the PGB, fostered consensus building between the countries and economies involved, and served as the interlocutor between the PGB and the international contractors.



The design and implementation of the surveys, within the framework established by the PISA Governing Board, is the responsibility of external contractors. For PISA 2015, the overall management of contractors and implementation was carried out by the Educational Testing Service (ETS) in the United States as the **Core 7** contractor. The OECD Secretariat worked closely with the International Project Director, Irwin Kirsch of ETS, to co-ordinate all aspects of implementation.

The additional tasks related to the implementation of PISA 2015 were carried out by six additional contractors – Cores 1 to 6.

Pearson in the United Kingdom developed the assessment frameworks as the **Core 1** contractor.

**Core 2** was led by ETS and focused on the development of the computer platform in co-operation with the Centre de Recherche Public Henri Tudor (CRP-HT) in Luxembourg.

**Core 3** focused on the instrument development, scaling and analysis and was led by ETS, with co-operation from cApStAn Linguistic Quality Control in Belgium for linguistic quality control, the University of Luxembourg, University of Heidelberg, GESIS and the Center for Educational Technology in Israel for test development, the Unité d'analyse des systèmes et des pratiques d'enseignement (aSPe) at the University of Liège in Belgium for coding training for open-constructed items, the International Association for Evaluation of Educational Achievement (IEA) in the Netherlands for the data management software, and HallStat SPRL in Belgium for the translation referee.

**Core 4** focused on Survey Operations and was implemented by Westat in the United States.

**Core 5** focused on sampling and was implemented by Westat in the United States in co-operation with the Australian Council for Educational Research (ACER).

**Core 6** focused on the questionnaire frameworks and questionnaire development and was carried out by the Deutsches Institut für Internationale Pädagogische Forschung (DIPF) in Germany, with the co-operation of Statistics Canada.

Annex G lists the staff and consultants associated with the core contractors who have made significant contributions to the development and implementation of the project.

## PISA 2015 PUBLICATIONS

This Technical Report is designed to describe the technical aspects of the project at a sufficient level of detail to enable review and, potentially, replication of the implemented procedures and technical solutions to problems. It therefore does not report the results of PISA 2015 which have been published in *PISA 2015 Results (Volume I): Excellence and Equity in Education* (OECD, 2016a) and *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools* (OECD, 2016b). Further results are reported in Volume III (OECD, 2017a), which discusses Students' Well-Being, Volume IV (OECD, 2017b), which reports on Students' Financial Literacy and Volume V (2017c), which delves into collaborative problem solving.

## References

- OECD** (2017a), *PISA 2015 Results (Volume III): Students' Well-Being*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264273856-en>.
- OECD** (2017b), *PISA 2015 Results (Volume IV): Students' Financial Literacy*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264270282-en>.
- OECD** (2017c), *PISA 2015 Results (Volume V): Collaborative Problem Solving*, OECD Publishing, Paris.
- OECD** (2017d), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264281820-en>.
- OECD** (2016a), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264266490-en>.
- OECD** (2016b), *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264267510-en>.
- OECD** (2014a), *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264208780-en>.
- OECD** (2014b), *PISA 2012 Results: Creative Problem Solving (Volume V): Students' Skills in Tackling Real-Life Problems*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264208070-en>.
- OECD** (2014c), *PISA 2012 Results: Students and Money (Volume VI): Financial Literacy Skills for the 21st Century*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264208094-en>.
- OECD** (2013a), *PISA 2012 Results: Excellence through Equity (Volume II): Giving Every Student the Chance to Succeed*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264201132-en>.
- OECD** (2013b), *PISA 2012 Results: Ready to Learn (Volume III): Students' Engagement, Drive and Self-beliefs*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264201170-en>.
- OECD** (2013c), *PISA 2012 Results: What Makes Schools Successful (Volume IV): Resources, Policies and Practices*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264201156-en>.
- OECD** (2011), *PISA 2009 Results: Students On Line: Digital Technologies and Performance (Volume VI)*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264112995-en>.
- OECD** (2010a), *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264091450-en>.
- OECD** (2010b), *PISA 2009 Results: Overcoming Social Background: Equity in Learning Opportunities and Outcomes (Volume II)*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264091504-en>.
- OECD** (2010c), *PISA 2009 Results: Learning to Learn: Student Engagement, Strategies and Practices (Volume III)*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264083943-en>.
- OECD** (2010d), *PISA 2009 Results: What Makes a School Successful?: Resources, Policies and Practices (Volume IV)*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264091559-en>.
- OECD** (2010e), *PISA 2009 Results: Learning Trends: Changes in Student Performance since 2000 (Volume V)*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264091580-en>.
- OECD** (2007), *PISA 2006: Science Competencies for Tomorrow's World*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264040014-en>.
- OECD** (2004), *Learning for Tomorrow's World: First Results from PISA 2003*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264006416-en>.
- OECD** (2001), *Knowledge and Skills for Life: First Results from PISA 2000*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264195905-en>.
- OECD/UNESCO Institute for Statistics** (2003), *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264102873-en>.
- Walker, Maurice** (2011), *PISA 2009 Plus Results: Performance of 15-year-olds in reading, mathematics and science for 10 additional participants*, ACER Press, Melbourne.



2

# Test design and test development

<b>Introduction .....</b>	30
<b>PISA 2015 integrated design .....</b>	30
<b>Overview of the field trial assessment design.....</b>	35
<b>Overview of the main survey assessment design.....</b>	36
<b>The 2015 assessment frameworks .....</b>	43
<b>Role of the subject matter expert groups in item development.....</b>	44
<b>PISA 2015 test development.....</b>	44
<b>Field trial .....</b>	49
<b>Main survey.....</b>	51



## INTRODUCTION

This chapter describes the assessment design for PISA 2015 as well as the processes used by the PISA Core 3 contractor, Educational Testing Service (ETS), and the international test development team to develop the tests for the 2015 cycle. Those tests included:

- science, the major domain in 2015
- reading and mathematics, the two minor domains
- collaborative problem solving (CPS), the innovative domain for this cycle
- financial literacy, an international option.

For the 2015 cycle, under the guidance of the PISA Governing Board (PGB), it was decided to move from a primarily paper-based delivery survey that included optional computer-based modules to a fully computer-delivered survey. A paper-based version of the assessment that included only trend units was developed for the small number of countries that did not implement the computer-based survey. The computer-based delivery mode allows PISA to measure new and expanded aspects of the domain constructs. In science, the addition of interactive tasks allowed students to manipulate variables in simulated scientific enquiries. Interactive chat-based tasks with branching based on student responses were used to assess collaborative problem solving.

Equally critical in 2015 was the introduction of an innovative assessment design that emphasised improved trend measurement and enhanced coverage of minor domains. The ability to establish and maintain trends over time is a goal for PISA that has been clearly and repeatedly articulated by the PGB and participating countries. For the first time in 2015, the integrated design for the assessment increased the number of items for the minor domains to previous major domain levels, reducing the potential for introducing systematic measurement error because of reduced domain coverage from one cycle to the next. Due to these changes, the design for PISA 2015 strengthened the measurement of trends, by helping to strengthen construct coverage for the minor domain cycles in PISA. It also reflected an innovative conceptual approach that looked at PISA from a broad perspective and focused on a nine-year survey cycle during which scientific, reading and mathematical literacy would each be assessed as a major domain.

## PISA 2015 INTEGRATED DESIGN

The goals for the integrated assessment design in PISA 2015 included:

- improving the measurement of trends over time across the three core PISA domains
- minimising respondent burden while maximising the range of information obtained for each domain assessed
- accurately describing the proficiencies of nationally representative samples of 15-year-olds in each country, including relevant subpopulations
- associating these proficiencies with a range of indicators in policy-relevant areas.

To meet these goals, the design for the assessment included a re-conceptualisation of the assessment of the minor domains that would diminish differences in domain coverage across cycles, a linking study to evaluate and control for potential mode effects when moving from a paper-based to a computer-based assessment, and computer administration as the primary mode of delivery for all core domains.

Among other things, this design increased the number of items, improving construct coverage for the minor domains, which then allowed for a new methodological approach to be employed. More importantly, the methodology implemented in 2015 incorporated all available data from previous cycles, up to the last major domain cycle, for scaling and analysis, providing a solid base for linking across cycles and between paper-based and computer-based administrations for all cognitive scales. This is in contrast to previous cycles where scaling was conducted for each cycle and then equated to previous results through a single transformation. Taken together, these design and methodological innovations served to improve comparability across countries, stabilise parameter estimations and the measurement of trends, and improve the reliability of inferences formed from the data.

### Minimising the distinction between major and minor domain coverage

Any assessment must contend with two types of errors – random and systematic. Random errors do not result in bias but do increase uncertainty and, therefore, affect only the precision of results. Systematic errors, on the other hand, introduce bias, especially in the measurement of trends, and are less desirable because their direction is unknown and not easily

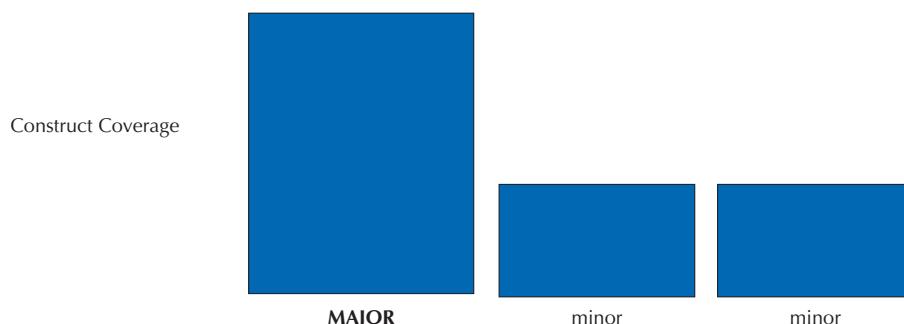


quantified or controlled for by statistical means. All large-scale surveys such as PISA, struggle with these two sources of error and aim to control them by optimising the assessment design, as well as sample size, sampling procedures, and other contributing factors. An increase in random errors reduces the ability to detect differences among groups of interest and can typically be offset by increasing sample size. However, an increase in systematic errors not only reduces the ability to detect differences, but also may lead to the attribution of false differences in size and direction; i.e., differences that are considered significant, even though the true differences are negligible, or even zero. Because of the possibility of introducing bias, a reduction in systematic errors is generally preferable over a reduction of random error components.

Figure 2.1 below illustrates the relative difference in construct coverage between the major and minor domains as implemented in PISA from 2000-2012. The vertical height of each bar represents the proportion of items measured in each assessment cycle by domain, while the width conveys the relative number of students who respond to each item within each domain. The reduced height of the bars for the minor domains represents the relative reduction in the number of items in that domain and therefore the degree to which construct coverage has been reduced.

■ Figure 2.1 ■

#### **Comparison of construct coverage in the 2000-2012 PISA design by major and minor domains**

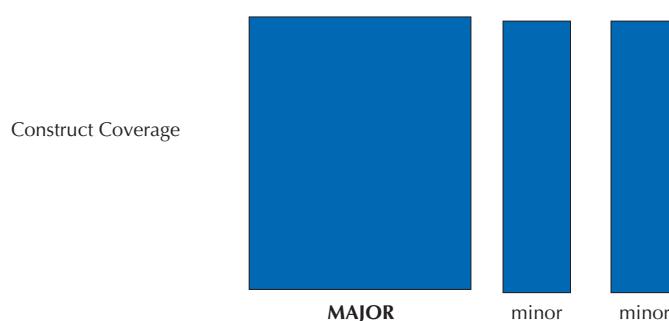


The new design used in PISA 2015 was intended to stabilise the trend and reduce potential systematic bias due to lack of domain coverage, by including more items in each minor domain than had been included in previous cycles, while reducing the number of students responding to each item. This strategy kept the volume of response data per student consistent across cycles, and increased the construct coverage for the minor domains, while reducing the number of students responding to each minor domain item per cycle. The result is that the construct representation for each minor domain is at a level comparable to the major domain cycle. As an added benefit, this approach reduces the potential for bias introduced due to item-by-country interactions in the subset of items that would have been selected for administration when the switch from major to minor domains in the previously used design occurred. This design both stabilises and improves the measurement of the minor domain, and its trend.

The approach adopted in 2015 is represented graphically in Figure 2.2 below. As represented by the height of the bars, the construct coverage for the minor domain is comparable to the major domain design, at the expense of reducing the number of students who respond to each of the minor domain trend items. This reduction of student responses per minor trend item is represented in the figure by the narrowing of the bars for the two minor domains.

■ Figure 2.2 ■

#### **Approach used to balance major/minor domains in 2015 and beyond**



Under this approach for measuring trends, each domain goes through a “domain rotation”, or a nine-year period that begins with a new or revised framework and continues with the two subsequent cycles in which it is a minor domain and then concludes with becoming a major domain once again. The end of the cycle involves another revision of the framework to reflect the current best thinking about assessment for the new major domain data collection. For example, as the major domain in 2015, the domain rotation for scientific literacy includes the 2015, 2018 and 2021 cycles with the next rotation beginning in 2024 when science will again be the major domain, with a newly revised framework. Thinking about designing the assessment in terms of this domain rotation clarifies the specific function of each cycle within that nine-year period, and the importance of maintaining the construct coverage in the minor cycles between two major domain cycles. Over a domain rotation, each major and minor cycle serves a specific function in terms of its contribution to the measurement of trends. Information about item functioning is carried across each domain rotation, with the choice of which items to carry forward being based on the most accurate item parameter estimation (occurring when a construct is measured as a major domain). The set of items that are carried forward in the rotation represents the full construct as covered in the initial major cycle, rather than a subset as in the prior minor domain design. In this way, the notion of a trend is defined both by the full coverage of the construct and by the statistical methodology employed.

To ensure trends are measured over longer periods of time, every time the framework for a major domain is revised – i.e. with the beginning of each domain rotation – a new set of items is developed to reflect the evolution of the construct. For PISA 2015, the revised framework for scientific literacy and the introduction of computer-based items broadened the construct beyond what was measured in 2006, the last time that scientific literacy was a major domain. This means that the PISA 2015 science scale must represent the revised framework while being linked to the existing scale represented by the previous framework through the set of existing trend items.

Linking proficiency scales in this way reduces the risk of introducing systematic errors in trend measures introduced by the new framework and item pool by establishing a point of connection between the backward-looking trend and forward-looking trend. Each updated construct is reflected by items that cover different aspects of the domain. Some items may reflect aspects unique to the old construct, most items will likely reflect aspects that are covered in both the old and revised construct, and there may be newly added items that reflect aspects introduced in the revised framework. This leads to the need to re-evaluate the combined set of items with respect to their relationship to the updated construct. Items that reflect both the old and revised framework will form the core of the combined scale, and items that are unique to either the old or revised framework will strengthen the link of this combined scale, looking backward to the old construct or forward to the new items added based on the revised framework. The generalised modelling framework allows the assignment of optimal weights to the items by re-estimating item parameters in each introductory cycle for the revised major domain. These optimal item weights facilitate the transition of the reported proficiency scale to the revised framework and the combined set of items, hence maintaining a link to prior assessments while transitioning to the new construct. Conceptualising the assessment design in this manner provides regular opportunities to introduce important and innovative ideas into (revised) major assessment domains. It also allows the opportunity to disentangle any changes in proficiency that result from differences in the construct and the way it is being measured.

### **Improving comparability and stabilising trends**

Establishing comparable and psychometrically sound scales requires design considerations as well as analytical choices that appropriately support this goal. The previous section explained several design innovations implemented to strengthen the comparability of results across countries and over assessment cycles. This section summarises a significant methodological shift that was introduced in 2015. In contrast to previous cycles, where scaling was conducted for each cycle and then equated to previous results through a single transformation, the methodology implemented in 2015 incorporated all available data for scaling and analysis, reaching back to the last introduction of the same domain as major domain, thus providing a solid base for linking across cycles and between paper- and computer-based administrations on all scales.

Equating scales refers to the process of transforming the scale scores of a more recent test onto the scale of a previous test form. Equating methods differ in terms of how they perform this transformation. In the most basic form of equating, a linear transformation is performed so that the main statistical properties of the transformed new test scores match those of the old test form. While there are equating methods for tests scored using classical test theory as well as for modern item response theory (IRT)-based tests, we focus on the latter here. In the context of *IRT equating*, the item parameters are typically estimated separately for both test forms and subsequently put on the same scale by means of a linear transformation. This approach can be mathematically shown to be inferior to so-called *IRT linking* that estimates item



parameters on the combined set of old and new data from the two or more test forms. The IRT linking approach provides a stronger equality constraint across parameters of the cycles to be linked through the items that are common to both test forms, while the linear IRT equating approach does not constrain the IRT model at all, but rather transforms indeterminate scales to match certain distributional moments. The assumptions made about the equality of item parameters can be tested statistically in this approach (e.g. Glas and Jehangir, 2013; Glas and Verhelst, 1995; Oliveri and von Davier, 2014, 2011). The IRT equating approach that only aligns average difficulty may implicitly assume parameter equality but typically does not involve this type of item level evaluation of parameter equality.

From 2000 to 2012, PISA relied on the IRT equating approach in which the anchor items common to the new and previous PISA cycles were used to find the transformation of the new data. This was carried out for each PISA cycle separately, so that over the first five cycles, four different transformations had to be used. This, in effect, produced five different sets of item parameters for those items that were used throughout the 2000–2012 cycles. In contrast, PISA 2015 introduced a comprehensive approach to scale linking in which all available data were combined to anchor the item parameters from the most recent PISA cycle together with data from past cycles. This was achieved by an IRT item calibration that ran across all PISA cycles and found common item parameters that maximised the fit of the IRT model to this comprehensive database. This linking approach utilised a common scale across all available data and represents the most rigorous and stable method of joining scales from different cycles. It preserved the inference structure of the proficiency scale by finding optimal item parameters for all items in the item pool, both for the common items that anchored the scale across cycles as well as items unique to a cycle. This approach generalised the methodologies utilised in other large-scale assessments (Mazzeo and von Davier, 2013) including, for example, the Programme for the International Assessment of Adult Competencies (PIAAC) that was jointly analysed and linked to the Adult Literacy and Life Skills Survey (ALL) and the International Adult Literacy Survey (IALS). The resulting item parameters can be transformed for all scales across all cycles in a way that maximally matches prior statistics for the assessment cycles that have been previously reported.

For illustration purposes, consider the PISA 2015 science domain. All data from 2015, when science was a major domain, were utilised to establish the forward-looking trend for 2018 and 2021. This included both the set of new items developed to represent the revised framework for science as well as the six clusters of trend items that were included in the main survey and for which additional data from 2006, 2009, and 2012 were used to link 2015 back to past cycles. This allowed the linking to have a positive impact on the comparability of results across countries, as one single set of parameters, instead of multiple sets, were used in the approach, and item parameter estimates based on multiple cycles have (after the appropriateness of parameter equality was tested) a smaller standard error. This also has a positive impact on the stability of trend measurements, since the best possible set of common parameters is found using this approach.

Let us for a moment assume that this was not true, that is a separate calibration in each cycle would provide the best possible link. In this case, the same argument would hold across countries within a cycle, so item parameters should be estimated by country, and each set of country-specific item parameters equated by aligning the average difficulty. Such an approach could lead to completely independent item estimates in each country and therefore would be neither appropriate nor acceptable because, for example, it would allow cases in which hard items in one country could be easy items in another. This would make comparisons across countries impossible.

The underlying assumption of linking and aligning scales is that (the vast majority of) items are comparable, and function the same in the sense of measurement invariance assumption (Meredith, 1993; Reise, Widaman and Pugh, 1993). This assumption is the basis for comparisons both across and within cycles across participating countries. If this were not the case, the PISA assessment would potentially measure something different in each country and in each cycle. It is for this reason that a multi-cycle scaling approach is used today by major large-scale assessments, including NAEP, TIMSS, PIRLS and now PISA.<sup>1</sup> Statistical modelling that combines multiple databases has a tradition also in other domains such as the analyses of psychological scales or data from patient reported outcomes. As noted by Curran et al. (2008) this type of integrative data analysis (IDA) has various advantages over separate statistical analyses that use post-hoc combination of estimates.

The approach used in PISA 2015 has several advantages. First, it produces more stable item parameter estimates since the item calibration takes place on a much larger database using IDA approaches. This is true both in terms of the item pool that is covering all previously used items in the nine-year cycle, as well as in terms of the sheer number of test takers within countries that contribute to the estimation of the parameters. In addition, the approach produces, with the addition of each cycle, a joint set of parameters that can be used moving forward. The set of parameters established in

2015 would be updated by the addition of the new major cycle in 2018 for reading (since new items are added through the renewal of a framework and major assessment domain) and could be kept fixed for the two minor cycles following a major cycle (as no new items are added), for example in science in 2018 and 2021. However, in other large-scale assessments it is common practice to adjust item parameter estimates by the addition of new data, but to keep the data from one or more previous cycles in the re-estimation. This is a basic principle behind statistical learning, either by keeping previously collected data and combining it with new data in the estimation, or by applying prior distribution in Bayesian estimation, which in effect does the same thing. The consistency of the estimated parameters across cycles is much higher under this approach than if item parameters are re-estimated each cycle independently.

Again, the comparison to country-specific scaling may make the point clearer. No consistency across countries would be assumed if item parameters were estimated separately by country and aligned post hoc by matching the means of difficulties. This approach of separate country specific estimation would not produce a link across participating countries; it merely aligns country-level parameters to a common average difficulty. This is an approach that would not be methodologically appropriate as parameters across countries and cycles are highly correlated (Oliveri and von Davier, 2014). Significantly different sets of parameters across countries would indicate a violation of measurement invariance (Meredith, 1993; Meredith and Teresi, 2006; Reise, Widaman and Pugh, 1993), so one central prerequisite of cross-country comparability would be violated. The same reasoning applies directly to the linking across PISA cycles. Therefore, the linking approach chosen for PISA 2015 follows an approach that utilises best practices to ensure measurement invariance through the invariance of item parameters across cycles and across participating countries.

## Goals and domain coverage

The design for the PISA 2015 core assessment was developed to provide participating countries with the following information:

- population distributions in science that reflect the new 2015 framework as well as links to the framework and scale developed in 2006
- population distributions in mathematics linked to the 2012 framework and mathematical literacy scale
- population distributions in reading linked to the 2009 framework and reading literacy scale
- population distributions in collaborative problem solving
- pairwise covariance estimates among each of the four domains
- three-way covariance information among the four cognitive domains including the three core PISA domains (reading, mathematics, and science)
- data to link the two modes of delivery: paper-and-pencil and computer-based.

In addition to the four core domains of science, mathematics, reading and collaborative problem solving, the PISA assessment included an optional assessment of financial literacy.

Figure 2.3 shows the number of clusters included in the PISA 2015 field trial and main study to meet the goals and coverage of the core domains assumed in this approach. As shown, all new items for science were developed as computer-based items. The design also included six clusters of trend items in science. There was no new item development for reading and mathematics in 2015, but the existing trend items in these domains were re-authored for the computer and delivered both in paper-and-pencil, and computer modes. Finally, collaborative problem-solving items were designed for administration only on the computer.

■ Figure 2.3 ■  
**Domain coverage for PISA 2015**

Domain	NEW (CBA only)		TREND (CBA and PBA)	
	Field trial	Main survey	Field trial	Main survey
Science	12 30-min clusters	6 30-min clusters	6 30-min clusters	6 30-min clusters
Reading			6 30-min clusters	6 30-min clusters
Mathematics			6 30-min clusters	6 30-min clusters
Collaborative problem solving	4 30-min clusters	3 30-min clusters		

Note: CBA stands for computer-based assessment and PBA, paper-based assessment.



## Studying mode effects in PISA 2015

One of the major goals for PISA 2015 was to ensure that trends could be maintained across paper- and computer-based modes of assessment. To that end, the PISA 2015 field trial included a mode effects study utilising methodologies that were adapted from experience with the OECD PIAAC study. Countries planning to use computer-based delivery in the main survey were required to include a within-school random sample of students taking paper-and-pencil forms in the field trial to test for mode effects and ensure trend measurement relative to performance in previous paper-based cycles.

## OVERVIEW OF THE FIELD TRIAL ASSESSMENT DESIGN

The field trial design needed to support several key goals including the evaluation of invariance of item parameters across previous PISA cycles and across the two modes for the 2015 cycle. In addition, initial item parameters needed to be estimated for the new science and collaborative problem-solving items. The computer-based assessment (CBA) included six intact trend clusters from science, reading and mathematics based on the assessment cycle when each was the major domain: 2006 for science, 2009 for reading and 2012 for mathematics. In order to test for mode effects, the design included a set of 18 paper-and-pencil forms covering the domains of reading, mathematics and science.<sup>2</sup> These were identical to the set of 18 computer-based test forms that consisted of items adapted and re-authored for computer administration. In addition, there were 12 test forms consisting of the new 2015 science tasks (forms 49-60 as shown below) and 12 new test forms combining those 2015 science items with the new collaborative problem solving tasks (forms 61-72). The schematic design illustrating the set of paper-and-pencil forms along with the set of CBA forms – including the CBA trend, CBA new science and CBA new science plus collaborative problem solving – is shown in Figure 2.4.

Note that, as shown in Figure 2.4, the field trial sample was 78 students in each of the 25 schools within each country. Of these students, 23% were assigned to Group 1 and took the trend test items on paper, 35% were assigned to Group 2 and took the trend test items on computer, and 42% were assigned to Group 3 and took the new science and CPS items on computer. Further sampling requirements for this design are discussed in Chapter 4.

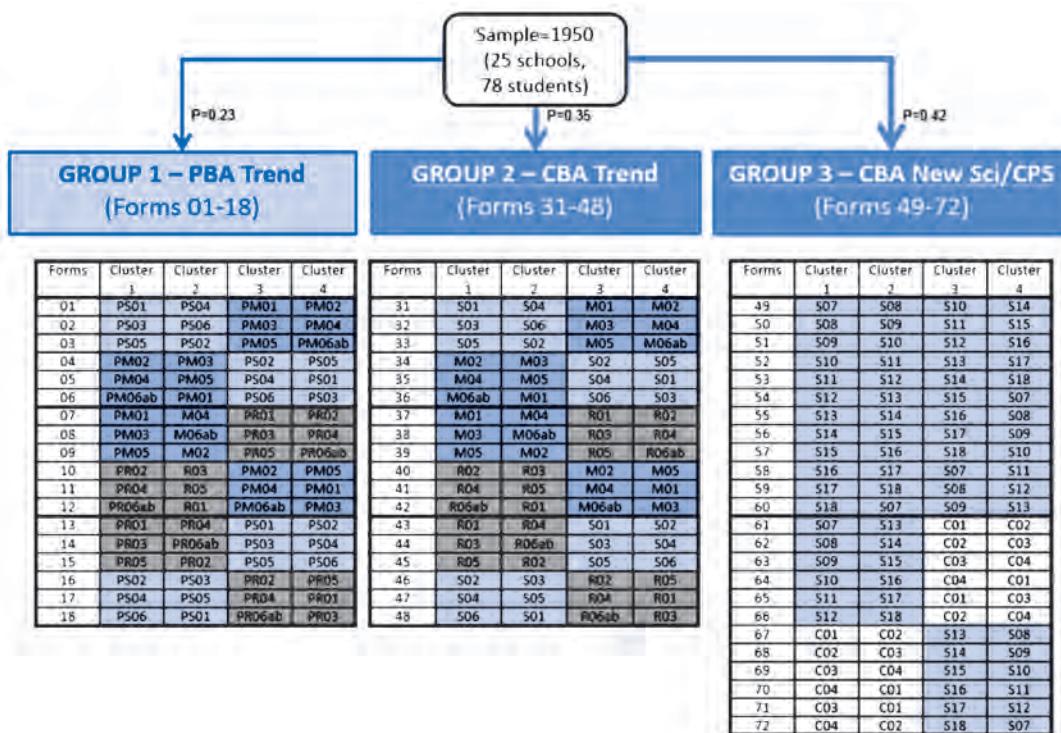
Where:

- *PR01-PR06* represent reading clusters in paper (trend)
- *PM01-PM06* represent mathematics clusters in paper (trend)
- *PS01-PS06* represent science clusters in paper (trend)
- *R01-R06* represent reading clusters in computer (trend)
- *M01-M06* represent mathematics clusters in computer (trend)
- *S01-S06* represent science clusters in computer (trend)
- *S07-S18* represent science clusters in computer (new)
- *C01-C04* represent collaborative problem-solving clusters in computer (new)
- *Subscripts a and b* are used to indicate standard (a) and easier (b) clusters, respectively.

Countries opting to deliver the paper-based version of the assessment in the main survey measured student performance with only paper-and-pencil forms in the field trial. Students were randomly assigned one of the 18 paper-and-pencil forms containing the trend items from two of the three core domains for PISA – reading (forms PR01-PR06), mathematics (forms PM01-PM06) and science (forms PS01-PS06).

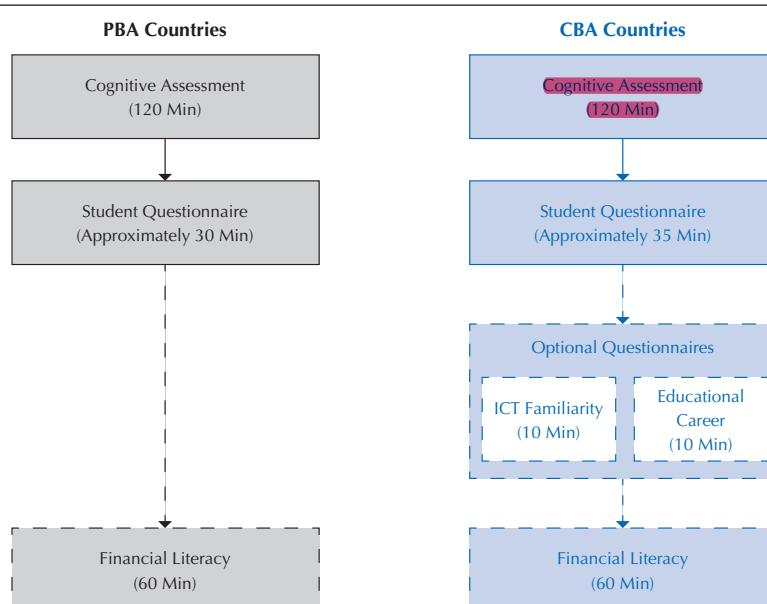
The findings of the field trial analyses on new and trend material in science, on the innovative domain of collaborative problem solving, and on the mode effect study are reported in Chapter 9.

■ Figure 2.4 ■

**Field trial computer-based assessment design, with collaborative problem solving****OVERVIEW OF THE MAIN SURVEY ASSESSMENT DESIGN**

The assessment design for PISA 2015 was planned so that the total testing time for measuring the four core domains of reading, mathematics, science and collaborative problem solving was two hours for each student. An overview of the flow of the integrated design for the PISA 2015 main survey is provided in Figure 2.5.

■ Figure 2.5 ■

**Overview of the PISA 2015 main survey integrated design\***

\* Note that while the optional assessment of financial literacy was offered for PBA countries and shown in Figure 2.5, none of the PBA countries in PISA 2015 opted to participate in this component.



### Paper-based integrated design

For PBA countries, the main survey tests included 30 forms. These are shown in Figure 2.6. All of the items included in the PBA test forms were taken from previous cycles of PISA. Each form included 1 hour of science items and items from at least one of the other two core domains. As a result, all students were administered science items, 56% of participating students were administered mathematics items, 56% reading items, and 12% were administered both reading and mathematics. The PBA was to be administered to 35 students in each of 150 schools. Further sampling requirements for this design are discussed in Chapter 4.

■ Figure 2.6 ■

#### Main survey paper-based assessment design

Percentage of students	Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
44%	1	PS01	PS02	PR01	PR02
	2	PS03	PS04	PR02	PR03
	3	PS05	PS06	PR03	PR04
	4	PS02	PS03	PR04	PR05
	5	PS04	PS05	PR05	PR06ab
	6	PS06	PS01	PR06ab	PR01
	7	PR01	PR03	PS01	PS02
	8	PR02	PR04	PS03	PS04
	9	PR03	PR05	PS05	PS06
	10	PR04	PR06ab	PS02	PS03
	11	PR05	PR01	PS04	PS05
	12	PR06ab	PR02	PS06	PS01
44%	13	PS01	PS03	PM01	PM02
	14	PS02	PS04	PM02	PM03
	15	PS03	PS05	PM03	PM04
	16	PS04	PS06	PM04	PM05
	17	PS05	PS01	PM05	PM06ab
	18	PS06	PS02	PM06ab	PM01
	19	PM01	PM03	PS01	PS03
	20	PM02	PM04	PS02	PS04
	21	PM03	PM05	PS03	PS05
	22	PM04	PM06ab	PS04	PS06
	23	PM05	PM01	PS05	PS01
	24	PM06ab	PM02	PS06	PS02
12%	25	PS01	PS02	PR01	PM01
	26	PS03	PS04	PM02	PR02
	27	PS05	PS06	PR03	PM03
	28	PM04	PR04	PS02	PS03
	29	PR05	PM05	PS04	PS05
	30	PM06ab	PR06ab	PS06	PS01

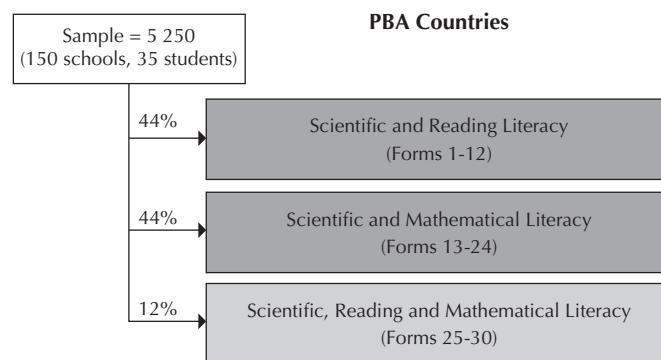
Where:

- PR01-PR06 represents reading clusters in paper (trend)
- PM01-PM06 represent mathematics clusters in paper (trend)
- PS01-PS06 represent science clusters in paper (trend)
- a and b represent standard clusters or easier clusters<sup>2</sup>, respectively.

Figure 2.7 presents a summary of the main survey PBA design. In the PBA design, 44% of students were assigned to one of 12 science and reading forms and another 44% were assigned to one of 12 science and mathematics forms. The remaining 12% of students were assigned to one of six science, reading and mathematics forms. This design included:

- 24 different test forms that combined two of the three domains, with 88% of students receiving one of these forms. In these forms, students took one hour of science plus one hour of another domain. These 24 forms provided strong pairwise covariance information between science and each of the two other domains.
- 6 additional forms that provided covariance information about the three domains. Twelve percent of students received one of these forms, which included one hour of science plus two 30-minute clusters from the minor domains.

■ Figure 2.7 ■  
**Main survey paper-based assessment design**



### **Computer-based integrated design**

For CBA countries including the collaborative problem-solving (CPS) assessment, the main survey included 66 forms (forms 31-96). These are shown in Figure 2.8. Under the full design, all sampled students responded to science items, 41% responded to mathematics items, 41% responded to reading items and 30% to CPS items. In addition, 4% responded to each possible combination of 2 of the minor domains.

For the five countries not participating in the collaborative problem-solving assessment, only 36 forms were included in the design (forms 31-66) and the percentages for this alternative design are also represented in Figure 2.8.

■ Figure 2.8 [Part 1/2] ■  
**Main Study Computer-Based Assessment Design**

Percentage of students	Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
33% (No CPS: 46%)	31	S	S	R01	R02
	32	S	S	R02	R03
	33	S	S	R03	R04
	34	S	S	R04	R05
	35	S	S	R05	R06ab
	36	S	S	R06ab	R01
	37	R01	R03	S	S
	38	R02	R04	S	S
	39	R03	R05	S	S
	40	R04	R06ab	S	S
	41	R05	R01	S	S
	42	R06ab	R02	S	S
33% (No CPS: 46%)	43	S	S	M01	M02
	44	S	S	M02	M03
	45	S	S	M03	M04
	46	S	S	M04	M05
	47	S	S	M05	M06ab
	48	S	S	M06ab	M01
	49	M01	M03	S	S
	50	M02	M04	S	S
	51	M03	M05	S	S
	52	M04	M06ab	S	S
	53	M05	M01	S	S
	54	M06ab	M02	S	S



■ Figure 2.8 [Part 2/2] ■

### Main Study Computer-Based Assessment Design

Percentage of students	Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
4% (No CPS: 8%)	55	S	S	M01	R01
	56	S	S	R02	M02
	57	S	S	M03	R03
	58	S	S	R04	M04
	59	S	S	M05	R05
	60	S	S	R06ab	M06ab
	61	R01	M01	S	S
	62	M02	R02	S	S
	63	R03	M03	S	S
	64	M04	R04	S	S
	65	R05	M05	S	S
	66	M06ab	R06ab	S	S
	67	S	S	C01	M01
	68	S	S	M02	C02
	69	S	S	C03	M03
4% (No CPS: NA)	70	S	S	M04	C03
	71	S	S	C02	M05
	72	S	S	M06ab	C01
	73	M01	C02	S	S
	74	C03	M02	S	S
	75	M03	C01	S	S
	76	C01	M04	S	S
	77	M05	C03	S	S
	78	C02	M06ab	S	S
	79	S	S	R01	C01
	80	S	S	C02	R02
	81	S	S	R03	C03
	82	S	S	C03	R04
	83	S	S	R05	C02
4% (No CPS: NA)	84	S	S	C01	R06ab
	85	C02	R01	S	S
	86	R02	C03	S	S
	87	C01	R03	S	S
	88	R04	C01	S	S
	89	C03	R05	S	S
	90	R06ab	C02	S	S
	91	S	S	C01	C02
	92	S	S	C02	C03
	93	S	S	C03	C01
	94	C02	C01	S	S
	95	C03	C02	S	S
	96	C01	C03	S	S

Where:

- R01-R06 represent reading clusters in computer (trend)
- M01-M06 represent mathematics clusters in computer (trend)
- S represents science clusters in computer (trend and new)
- C01-C03 represent CPS clusters in computer (new)
- a represents standard clusters and b represents easier clusters.

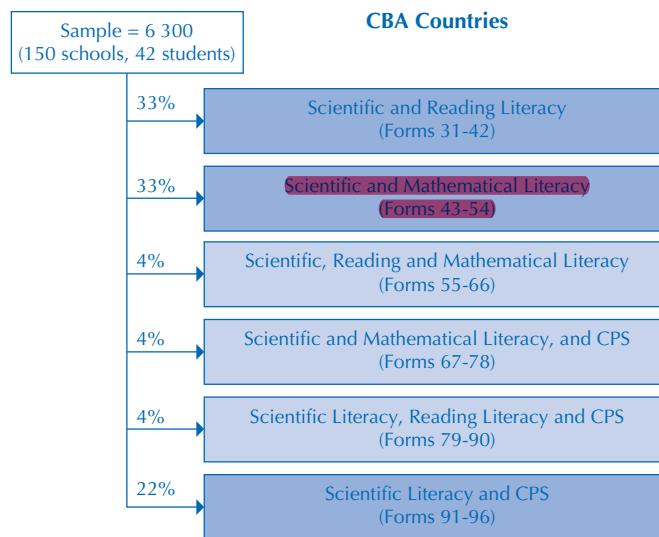
Figure 2.9 presents a summary of the main survey computer-based assessment design which was to be administered to 42 students in each of the 150 schools within each country. The design included:

- 30 different test forms that combined two of the four domains, with 88% of students receiving one of these forms. In these forms, students took one hour of science plus one hour of another domain. These 30 forms provided strong pairwise covariance information between science and each of the three other domains.
- 36 additional forms provided covariance information among the three minor domains. Twelve percent of students received one of these forms, which included one hour of science plus two 30-minute clusters from two of the other three domains.

Further sampling requirements for this design are discussed in Chapter 4.

■ Figure 2.9 ■

### Main survey computer-based assessment design



The rotation of clusters identified the form assigned to each student. This cluster rotation was determined by a multi-step random process that occurred at the time students were sampled. This process, described in more detail in the following section, was only possible because of the computer-delivered testing environment used in PISA 2015.

### Main study form assignment for the computer-based assessment

The rotation of clusters – which identified the form to be received by each student – occurred in a multistep process when students were sampled. KeyQuest, the sampling software used in PISA 2015, assigned two random numbers to each sampled student.

- CC was a two-digit random number that represented the base form for the test (i.e., 31-96 for regular students or 99 for UH students – see “UH Form” section for more information). This number met the probability constraints described for the CBA forms.
- S was a one-digit random number that was used as a lookup number to select the two science clusters that would be inserted into the base form of the test. This number was between 1 and 6, inclusive, and was uniformly distributed.

These random numbers were encoded into the login information for the computer platform that was assigned by KeyQuest.

#### **STEP 1: Assignment of the base test form**

The first step was assigning base test forms. This assignment was based on the two-digit random number identified as “CC”. This number ranged from 31-96 and was directly linked to a specific base test form as shown in Figure 2.8. These base test forms identified the actual location and clusters for mathematics, reading and CPS, but only identified the location of science, not the specific clusters – the specific science clusters were not assigned until Step 2 and therefore were only identified as “S” at this point. The probability of assignment of each form type varied from 33% to 4% as shown in Figure 2.8.

For countries not participating in the assessment of CPS, the two-digit random number ranged from 31-66, representing the forms without CPS. The probability of assignment of form also changed. For non-CPS countries, 46% of students were assigned forms 31-42 and 46% were assigned forms 43-54, while 8% were assigned forms 55-66. In other words, 92% of students received a form that consisted of four 30-minute clusters assembled from two domains. These percentages are shown in brackets in the first column of Figure 2.8.

#### **STEP 2: Assignment of science Clusters**

The second step was the assignment of science clusters. There were 36 possible science cluster combinations, with clusters S1 – S12 rotating as shown in Figure 2.10. Combinations 1-18 included both trend and new clusters; 19-33 included only new clusters; and 34-36 included only trend clusters.



■ Figure 2.10 ■

### Main study computer-based assessment combinations of science clusters

Science cluster combination		
N	S	S
1	S01	S07
2	S01	S10
3	S02	S08
4	S03	S09
5	S03	S12
6	S04	S07
7	S04	S10
8	S05	S11
9	S06	S12
10	S07	S06
11	S08	S01
12	S08	S05
13	S09	S02
14	S09	S06
15	S10	S03
16	S11	S02
17	S11	S04
18	S12	S05

Science cluster combination		
N	S	S
19	S07	S08
20	S07	S09
21	S07	S11
22	S08	S10
23	S08	S12
24	S09	S08
25	S09	S11
26	S10	S07
27	S10	S09
28	S10	S12
29	S11	S08
30	S11	S10
31	S12	S07
32	S12	S09
33	S12	S11
34	S02	S04
35	S05	S01
36	S06	S03

■ Figure 2.11 ■

### Lookup table for random number "S": Assignment of science cluster combinations

Base form (CC)	Random number (S)					
	1	2	3	4	5	6
31	1	13	6	9	22	25
32	2	16	12	10	31	32
33	11	5	17	14	26	29
34	35	4	7	19	23	30
35	34	15	8	20	24	28
36	3	36	18	21	27	33
37	35	4	7	19	23	30
38	34	15	8	20	24	28
39	3	36	18	21	27	33
40	1	13	6	9	22	25
41	2	16	12	10	31	32
42	11	5	17	14	26	29
43	1	13	6	9	22	25
44	2	16	12	10	31	32
45	11	5	17	14	26	29
46	35	4	7	19	23	30
47	34	15	8	20	24	28
48	3	36	18	21	27	33
49	35	4	7	19	23	30
50	34	15	8	20	24	28
51	3	36	18	21	27	33
52	1	13	6	9	22	25
53	2	16	12	10	31	32
54	11	5	17	14	26	29
55	1	13	6	9	22	25
56	2	16	12	10	31	32
57	11	5	17	14	26	29
58	35	4	7	19	23	30
59	34	15	8	20	24	28
60	3	36	18	21	27	33
61	35	4	7	19	23	30
62	34	15	8	20	24	28
63	3	36	18	21	27	33

Base form (CC)	Random number (S)					
	1	2	3	4	5	6
64	1	13	6	9	22	25
65	2	16	12	10	31	32
66	11	5	17	14	26	29
67	1	13	6	9	22	25
68	2	16	12	10	31	32
69	11	5	17	14	26	29
70	35	4	7	19	23	30
71	34	15	8	20	24	28
72	3	36	18	21	27	33
73	35	4	7	19	23	30
74	34	15	8	20	24	28
75	3	36	18	21	27	33
76	1	13	6	9	22	25
77	2	16	12	10	31	32
78	11	5	17	14	26	29
79	1	13	6	9	22	25
80	2	16	12	10	31	32
81	11	5	17	14	26	29
82	35	4	7	19	23	30
83	34	15	8	20	24	28
84	3	36	18	21	27	33
85	35	4	7	19	23	30
86	34	15	8	20	24	28
87	3	36	18	21	27	33
88	1	13	6	9	22	25
89	2	16	12	10	31	32
90	11	5	17	14	26	29
91	1	13	6	9	22	25
92	2	16	12	10	31	32
93	11	5	17	14	26	29
94	35	4	7	19	23	30
95	34	15	8	20	24	28
96	3	36	18	21	27	33

The assignment of these combinations of science clusters was based on the one-digit random number "S". This number ranged from 1-6<sup>3</sup>, was uniformly distributed, and was used in combination with the base form (e.g., selected by the first two-digit random number) to identify which combination of science clusters a student received. Figure 2.11 shows the lookup table where the 31-96 base forms were identified by the rows and the 1-6 lookup numbers are identified by the columns. The combination of these two numbers was used to identify which of the 36 possible combinations of science clusters was used with the assigned base test form.

As an example of how this assignment process worked, suppose a student was assigned random numbers of CC = 37 and S = 4. Based on this information, the assignment of cognitive clusters was: i) base test form 37 which included two reading clusters (R01 and R03) and two science clusters; and ii) lookup number 4 that identified science cluster combination 19, which included science clusters S07 and S08. As a result, this student received a test composed of the following clusters:

Cluster 1	Cluster 2	Cluster 3	Cluster 4
R01	R03	S07	S08

### Une heure (UH) form

Consistent with previous cycles, a special one-hour test, referred to as the "Une Heure" (UH) form, was prepared for students with special needs. The selected items were among the easier items in each domain and had a more limited reading load. The UH form contained about half as many items as the other instruments, with each cluster including from seven to nine items. The UH form was comprised of about 50% science, 25% mathematics and 25% reading items.

The UH form included two clusters of science (SU1 and SU2), one cluster of reading (RU1), and one cluster of mathematics (MU1). The assignment of this booklet followed the approach described previously for the assignment of the base test form. The UH form was assigned base form 99 (as shown in Figure 2.12) and the two-digit random number, was not considered for selection of this form.

■ Figure 2.12 ■

#### Main survey UH form design

Form	Cluster 1	Cluster 2	Cluster 3	Cluster 4
99(UH)	SU1	SU2	RU1	MU1

The UH form was accompanied by a UH student questionnaire that included a subset of items from the regular questionnaire (primarily trend items) in a single form design that was administered in CBA only, as no PBA countries chose to administer the UH Form.

### Assessment of financial literacy

The assessment of financial literacy was offered as an international option in PISA 2015. It was based on a slightly re-ordered version of the items from PISA 2012 and included all but the one released item from 2012 with four new items added. In the main survey, financial literacy was available only as a computer-based assessment because countries participating in this option were all CBA countries. It was administered to a subsample of the PISA sample that took combinations of mathematics, reading and science items.

Countries opting for the financial literacy assessment were required to participate in the mode effect study and administer paper and computer versions of instruments in the field trial. The approach for the field trial included administration of financial literacy forms to a subsample of the PISA sample that took combinations of mathematics and reading items.

For the field trial design the following two groups also took financial literacy:

- Group 1 (PBA trend) included students taking Booklets 07-12 (reading and mathematics). Within each school there were approximately six students taking these booklets, all of whom also took financial literacy. This group took financial literacy as a paper instrument.
- Group 2 (CBA trend) included forms 37-42 (reading and mathematics). Within each school there were approximately nine students taking these forms, with all students also taking financial literacy. This group took financial literacy as a computer instrument.

This design provided a field trial sample size of approximately 375 students per country, with about 150 students taking the paper version, and 225 students taking the computer version.

For the main survey, the assessment instruments included 43 items, of which 39 were trend items and 4 were new items. These items were organized into two 30-min clusters that were rotated into two forms with each student taking both



clusters. The approach for the main study included the administration of financial literacy forms to a subsample of the PISA sample that took the core domains.

Students selected to take financial literacy were a subgroup of the students sampled based on the form they were assigned for the assessment of the core domains. The following forms were selected:

- forms 31, 33, 39 and 42 (science and reading): about 693 students per country
- forms 43, 45, 51 and 54 (science and mathematics): about 693 students per country
- forms 55-66 (science, mathematics and reading): about 252 students per country.

In total about 11 students in each school were subsampled for financial literacy, resulting in a total sample of approximately 1,650 students per country. This was the case for all CBA countries, including those few who took financial literacy but not CPS.

## THE 2015 ASSESSMENT FRAMEWORKS

For each PISA domain, an assessment framework is produced to guide instrument development and interpretation in accordance with the policy requirements of the PISA Governing Board. The frameworks define the domains, describe the scope of the assessment, specify the structure of the test – including item format and the preferred distribution of items according to important framework variables – and outline the possibilities for reporting results. For PISA 2015, subject matter expert groups (SMEGs) were convened by the Core 1 contractor to develop frameworks for science and collaborative problem solving.<sup>4</sup> The reading and mathematics frameworks were based on those developed for the 2009 and 2012 assessment cycles, respectively, when these domains were treated as major domains.

### Science

The 2015 framework for science emphasises the importance of educating all young people to become informed, critical users of scientific knowledge. To understand and engage in critical discussion about issues that involve science and technology requires three domain-specific competences: knowledge of the fundamental ideas of science and the questions that frame the practice and goals of science, knowledge and understanding of scientific enquiry, and the ability to interpret data and evidence scientifically. Thus, the 2015 framework defines science as follows:

*Science is the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen.*

*A scientifically literate person, therefore, is willing to engage in reasoned discourse about science and technology which requires the competencies to:*

- **Explain phenomena scientifically** – recognise, offer and evaluate explanations for a range of natural and technological phenomena.
- **Evaluate and design scientific enquiry** – describe and appraise scientific investigations and propose ways of addressing questions scientifically.
- **Interpret data and evidence scientifically** – analyse and evaluate data, claims and arguments in a variety of representations and draw appropriate scientific conclusions.

The assessment tasks focused on three dimensions of science:

- *competencies*, including explaining phenomena scientifically, evaluating and designing scientific enquiry, and interpreting data and evidence scientifically, as described above
- *knowledge*, including knowledge of both the natural world and technological artefacts (content knowledge), knowledge of how such ideas are produced (procedural knowledge), and an understanding of the underlying rationale for these procedures and the justification for their use (epistemic knowledge)
- *contexts*, including personal, local/national and global issues.

### Collaborative problem solving

As the innovative domain in the 2015 cycle, the collaborative problem solving assessment focuses on skills that have become increasingly important both across educational settings and in the workforce. The domain is defined as follows:

*Collaborative problem solving is the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution.*

This definition incorporates three core collaborative problem solving competencies: establishing and maintaining shared understanding; taking appropriate action to solve the problem; and establishing and maintaining team organisation. Additionally, the collaborative problem solving framework incorporated the four problem solving processes included in the PISA 2012 problem solving framework: exploring and understanding; representing and formulating; planning and executing; monitoring and reflecting. The three major CPS competencies were crossed with the four major individual problem solving processes forming a matrix of specific skills to be assessed in PISA 2015. As shown in Figure 2.13, this identified the dimensions of the tasks developed for the collaborative problem solving domain.

■ Figure 2.13 ■

### Matrix of collaborative problem solving skills for PISA 2015

	(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organisation
(A) Exploring and Understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve problem
(B) Representing and Formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organisation (communication protocol/rules of engagement)
(C) Planning and Executing	(C1) Communicating with team members about the actions to be/ being performed	(C2) Enacting plans	(C3) Following rules of engagement, (e.g., prompting other team members to perform their tasks)
(D) Monitoring and Reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organisation and roles

### ROLE OF THE SUBJECT MATTER EXPERT GROUPS IN ITEM DEVELOPMENT

As the contractor for instrument development, Core 3 was responsible for working with the subject matter experts in all domains. The proposed selection of trend items in the 2015 minor domains of reading and mathematics was shared with the subject matter expert groups (SMEGs) in September 2012. Proposals for adaptations to enable the display of longer texts in the computer-based reading units, along with a limited number of response mode adaptations in both domains, were shared with the subject matter experts for their input.

Core 3 worked with the expert groups for science and collaborative problem solving to understand their vision for the range and types of items to be developed for PISA 2015. To facilitate the transition from the work of Core 1 (framework development) to the instrument development activities, Core 3 retained the SMEG members who began work on the frameworks in early 2012. Core 3's work with the SMEGs began in June 2012 and focused on the following tasks:

- describing the kinds of items needed to assess the skills and abilities in each domain as those were defined in the framework
- reviewing and understanding the proposed assessment design in order to define the number and types of items that were needed for each of the domains
- defining the behaviours of interest for the computer-based tasks
- defining the intersection between the kinds of functionality that might be desirable for measuring the constructs and the functionality that was practicable to implement in the assessment.

Work with the subject matter experts continued beyond the initial meetings through instrument development and data analysis. For science and collaborative problem solving, SMEG members played an important role in reviewing assessment tasks as they were developed, providing input into the analysis of the field trial data, approving the set of items for the main survey, and working with development and analysis staff to develop the described scales used for reporting the PISA 2015 results.

### PISA 2015 TEST DEVELOPMENT

Test development for the PISA 2015 cycle began in mid-2012. The transition to a computer-delivered assessment, along with the new assessment design for this cycle that required many more trend items than had been used in past cycles, resulted in a number of development challenges that were unique to this cycle. In addition, the number of science items developed and field tested was much larger than usual for a major domain to allow for the possibility of an adaptive design in the main survey – an option which, in the end, was not implemented in this cycle.



## Computer-based assessment: Screen design and interface

A critical step in the item development process for PISA 2015 was creating a screen design that would be forward looking while still ensuring that PISA could continue to provide reliable trend data. This meant the design needed to support the range of display options and interaction modes required by new, innovative items while also facilitating the display of paper-and-pencil trend items being moved to the computer for reading, mathematics, science and financial literacy. An equally important consideration was the impact of the screen design across the range of languages in participating countries.

Given these considerations, Core 3 proposed a vertically split screen design in which the stimulus would be displayed in a pane on the left and the question or task in a pane on the right.<sup>5</sup> The panes were adjustable in width to accommodate varying content and, where appropriate, a single-pane design was also used. The vertically split design achieved a number of important goals in that it:

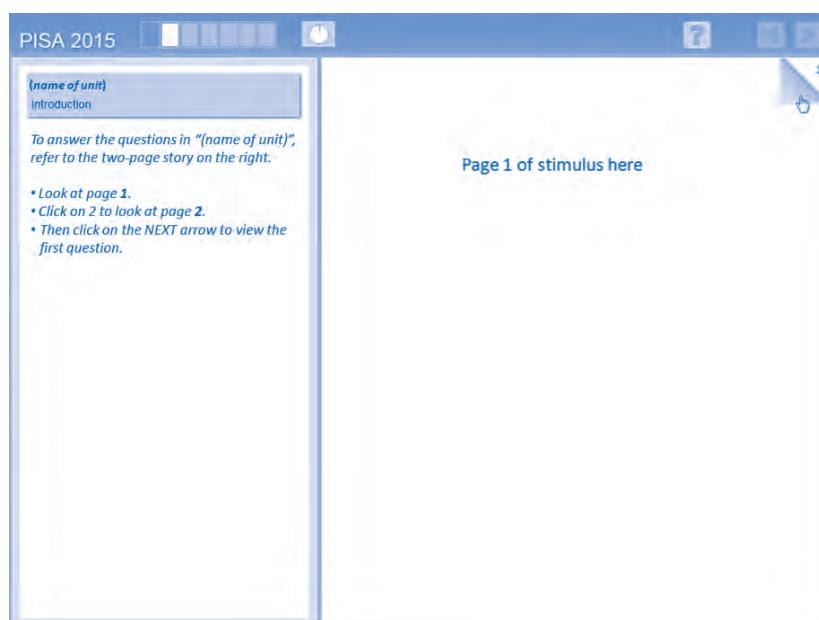
- facilitated the display of paper-and-pencil trend items that were moved to computer delivery
- allowed text to be formatted with shorter line lengths, improving readability
- accommodated displays across a variety of languages
- allowed PISA to take advantage of wider computer screens that are likely to become more prevalent in the future.

A paper outlining the proposed screen design for the PISA 2015 cognitive instruments was submitted to the OECD Secretariat on 26 July 2012. In addition, an overview of the design was presented for discussion at the September 2012 Subject Matter Expert Group meetings for science, reading, mathematics and collaborative problem solving and at the meeting of the National Project Managers (NPMs) that same month. In cooperation with the OECD Secretariat, a revised version of the paper was submitted on 1 October 2012 as a background document for the October 2012 PGB meeting, where the design was formally approved.

### **Multi-page stimulus materials**

A number of stimulus materials, particularly in reading, were presented on more than a single page in the paper-based format and, similarly, occupied more than a single screen on the computer. After consultation with members of the Reading Expert Group, the decision was made to present longer texts on static screens with a paging interface that allowed students to move from page to page throughout the text. Of the 29 units included in the 2015 assessment, 66% were presented on a single screen, 31% required two screens and just one unit required three screens. Decisions about where to split the text across pages were driven by the need to keep the presentation as similar as possible to the paper-based display and to ensure that all languages would have the same information displayed on each page. Figure 2.14 shows the paging display used in PISA 2015.

■ Figure 2.14 ■  
**Paging navigation used in PISA 2015**



A number of safeguards were included to ensure that students saw all the pages in each unit and understood how to navigate among them.

- Students were introduced to the paging interface in the orientation.
- Prior to encountering the first question for any stimulus that spanned more than a single screen, students were instructed to click on each page of the stimulus, as shown in the directions on the left pane in Figure 2.14.
- The “NEXT” button did not become active until students had clicked on each page. Thus students could not proceed to the first question in the unit until they had viewed each page in the stimulus.
- Each turned down page corner was animated so it moved when students hovered the cursor over it. This animation was included to further draw students’ attention to the paging display.

### ***Navigation***

Decisions about how students would be allowed to navigate through the items also needed to be built into the interface design for PISA 2015. For the majority of units, students were able to move back and forth among items *within* a unit. They were not, however, able to move back and forth *among* units. Once students clicked on the “NEXT” button on the final item in a unit, a dialog box displayed a warning that the student was about to move on to the next unit and it would not be possible to return to previous items. At this point, students could either confirm that they wanted to go on or cancel the action and continue with the unit on which they had been working.

Navigation for the interactive science and collaborative problem solving items followed a somewhat different model in that students were not able to go back to a previous item within a unit. The branching within the chat-based interface for collaborative problem solving meant that students could not change their chat choices once they clicked on the “Send” button. Similarly, students were not able to rerun the simulated experiments associated with each item in a unit because this would make the log files for these items unduly complex. Both the CPS and science orientations introduced this navigation to students. In addition, a dialogue box following each item required that students confirm they were ready to continue to the next question.

### ***Response modes***

Across all domains, PISA 2015 included items requiring one of five different response modes:

- click on a choice
  - single-selection multiple choice (includes chat format)
  - multiple-selection multiple choice (click on one or more responses)
  - complex multiple choice (table with statements and a number of yes/no or true/false options)
  - click on an image
- numeric entry (only numbers, comma, period, dash and backslash could be entered)
- text entry (within a scrolling text box that did not constrain the length of a student response – consistent with what was possible for paper-and-pencil items)
- select from a drop-down menu
- drag and drop (including use of a slider).

### ***Orientations***

A general orientation introduced students to the screen design and those response modes that were common across most domains. Students received this orientation before beginning the test. Prior to beginning each section of the test, students received a very short domain-specific orientation with instructions specific to the domain in that section. For example, before beginning the reading section of the assessment, students were introduced to the paging interface for the longer stimulus materials.

### ***Trend items***

The assessment design for PISA 2015 required that six 30-minute clusters of trend items be taken from previous cycles for reading, mathematics and science. The number of items required to meet this design meant that all available existing items (e.g., items that had not been released in previous cycles) needed to be adapted for the computer and included in the field trial. All 83 of the unreleased 2012 mathematics items were included in the PISA 2015 field trial.<sup>6</sup> In reading,



44 of the items used in the 2012 cycle were used, along with 59 additional items taken from the 2009 cycle. For the science trend, 53 of the items included in the 2012 cycle were used, along with 30 items from the 2006 cycle and eight items from the 2003 cycle.<sup>7</sup> In total, the PISA 2015 field trial included 83 mathematics items, 104 reading items and 91 trend science items.

In general, the goal in adapting the trend items from a paper-based to computer-based assessment was to maintain the presentation of information and cognitive demands, in order to maintain trend measurement. The computer version of each trend item was mocked up in several languages to determine where adaptations might be required to ensure a consistent display. For example, with longer stimuli, it would not be acceptable to have information required to answer a question on the first screen in some languages but on the second screen in others, as that would be likely to affect item difficulty. The specific considerations for re-authoring and adaptations differed somewhat across domains.

For the trend reading items, the primary challenge was the presentation of longer and more complex stimuli. Of the 29 unique stimuli, 14 fit on screen with no adaptations, 10 were presented on two pages in the paper booklets and could be similarly presented on two screens using the paging interface previously described, and 6 required adaptations including a minor reduction in the size of images or displaying text on two screens where it had been on a single page in paper.

Display of the stimulus materials was not an issue for mathematics as these tended to be brief and fit well on the screen across languages. To allow students to show how they found an answer or, in a few cases, enter a formula where one was required as a response, the mathematics test included a tool called the equation editor which included a set of mathematical symbols unavailable on the standard keyboard. Students were taught how to use the tool in the orientation presented just prior to beginning the mathematics section of the assessment.

Several of the science trend units included multiple stimuli that were associated with different items. For example, the first item in the paper-based version would require students to read a short text, the second item would include a graph related to the same topic and the third would be associated with a table. In the computer-based version of such units, it was important to ensure that students noticed the new information that was displayed with each item. This was accomplished by changing the headings or titles displayed on the right side of the screen with each stimulus as well as changing the user instructions for each item to direct students to refer to that information.

Finally, the financial literacy trend items were moved quite seamlessly from paper to computer, requiring no stimulus adaptations or changes in response modes.

## New Items

To meet the expanded design for PISA 2015, six 30-minute clusters of new items were developed for science and four 30-minute clusters for collaborative problem solving. In total, 213 science items were developed and included in the field trial.<sup>8</sup> This set included 158 standard items embedded within 40 units and 55 interactive items associated with 10 units. The collaborative problem solving domain included seven units in the field trial with 187 associated score points. Finally, ten new items were developed for financial literacy, four of which were taken forward to the field trial.

## International test development team

Test development efforts were coordinated by Core 3 at ETS. As is the case with any large-scale international survey, it is important that the pool of tasks used in PISA reflect the range of contexts and experiences of students across participating countries. One way to meet this goal is by convening an international team of item developers. For PISA 2015, the international test development team included individuals from the Centre for Educational Technology in Tel-Aviv, Israel, the University of Luxembourg, and the GESIS-Leibniz Institute for the Social Sciences in Mannheim, Germany. These groups worked with submissions from 23 countries in science and seven in collaborative problem solving to develop the pool of items included in the PISA 2015 field trial.

## National submissions and reviews

A second method for ensuring that the item pool reflects the international context of an assessment such as PISA is to solicit item submissions from participating countries. Given the extremely tight development timeline for PISA 2015, Core 3 submitted a request for early submissions of stimuli and context ideas to the OECD Secretariat in July 2012. Those were shared by the OECD Secretariat with countries in August and resulted in a number of submissions prior to the first meeting of National Project Managers (NPMs) in September 2012. More detailed item submission guidelines were prepared for countries and distributed as documents for that meeting in September.



For science, submissions were organised in two rounds.

- In Round 1, which ended on November 1, 2012, countries were asked to submit sample contexts and ideas for interactive units. These materials were needed early in the development cycle as the interactive units required more time to design, program and test. Submissions for the non-interactive, or “standard” units, were encouraged in this round as well. Four countries submitted ideas for 13 interactive science units. In addition, 6 countries submitted 7 standard science units along with contexts for an additional 4.
- In Round 2, countries were asked to submit standard units only. These units could be accepted later in the process as they could be prepared for review more quickly. National Centres were asked to submit Round 2 items by mid-December 2012 so those items could be integrated into the country review cycle, allowing all participating countries to review the materials proposed for the field trial. In total, 23 countries submitted science units during this round.

Given the innovative nature of the collaborative problem solving domain, countries were asked to contribute to the item development process by submitting sample contexts and problem situations, or “abstracts”, to better ensure that the pool of CPS tasks reflected the cultural diversity across participating countries. An abstract submission form was developed to guide this process. Submissions were requested by November 1, 2012. Seven countries submitted CPS materials for consideration.

Countries had the opportunity to review and provide feedback on units developed by the international test development consortium and participating countries at three points during the assessment development process. Reviews were organised into two-week periods scheduled from late October 2012 to mid-January 2013, with each period focusing on different batches of items. Twenty-nine countries submitted reviews of the science items during the first review period, 40 during the second and 44 during the third. Content for collaborative problem solving was released in the form of abstracts for the first review. Feedback was provided by 27 countries. Detailed unit overviews with screen captures and descriptions of possible student actions were released for the second and third review periods, with 33 countries participating in the former and 38 in the latter.

Countries were also able to review the trend materials as computer-based units. Screen images of the reading and mathematics trend items were released during the first review period in October 2012 and the science trend units were released in Round 2.

### **Additional item reviews**

Newly developed units were submitted for translatability review at the same time they were released for country review.<sup>9</sup> Linguists representing different language groups provided feedback on potential translation, adaptation and cultural issues arising from the initial wording of items. Experts at cApStAn and the translation referee for the 2015 cycle were able to alert item developers to both general wording patterns and specific item wording that would be problematic for some translations and to provide suggested alternatives. This allowed item developers to make wording revisions at an early stage, in some cases simply using the alternatives provided and in others working with cApStAn to explore other possibilities.

Preparation of the French source version for all the tests’ units provided another opportunity to identify issues with the English source version related to content and expression that needed to be addressed. Development of the two source versions helped ensure that items were as culturally neutral as possible, identified instances where wording could be modified to simplify translation into other languages, and specified where translation notes would be needed to ensure the required accuracy in translating items to other languages.

In addition, user testing was conducted with students in both the United States and Luxembourg to identify where instructions might be improved or the interface reconsidered. The testing in Luxembourg was conducted with ten students and included seven units: two reading units that employed the paging interface, three mathematics units, each of which required students to use the equation editor tool and/or show their work, and two standard science units, which included the single-selection multiple choice, multiple-selection multiple choice, drag and drop, and type item types. The testing at ETS involved eight participants who were asked to work on one collaborative problem solving unit, one interactive science unit, a mathematics unit that included the equation editor and one reading unit that required the paging interface.

Information from these sessions was used to make revisions to one interface element in mathematics and correct several identified bugs. Equally important, the questions raised by study participants informed the development of the domain orientations, identifying areas where students needed instruction and practice before working on the assessment items.



### **Selection of new items for the field trial**

The 2015 item development process resulted in a total of 289 new science items: 231 standard items across 55 units and 58 interactive units across 11 units. Ten collaborative problem solving units were developed. Items were selected for inclusion in the field trial based on country reviews, feedback from the expert group and the distribution of items across the key categories as defined in the framework. Of the 213 selected science items, 65 percent, or a total of 140 items, originated from the national submissions received from 15 countries.

### **FIELD TRIAL**

The PISA 2015 field trial data collection timeline began in March 2014 and extended through August 2014 with 74 participating countries or economies across some 100 language versions. Countries moving to the computer-based assessment used both the computer-based and paper-based tests in the field trial in order to support the mode study for the trend items. The field trial tests for those countries testing solely in paper consisted of paper-based tests including only trend items from previous cycles. Assessment materials were prepared and released based on the field trial testing dates for each country.

### **Preparation of field trial instruments**

As part of the quality control procedures for PISA 2015, the Core 3 contractors assumed responsibility for migrating existing paper-based versions of the selected trend items to the computer for all computer-based countries. Core 3 also prepared all paper booklets used in the field trial for both paper- and computer-based countries. Countries were responsible for translating all new material and performing both linguistic and layout quality control checks for trend and new items in both modes. Where countries identified errors as a result of those checks, they were shared with the contractors who made any agreed-upon corrections.

#### **Computer-based trend items**

For countries with existing translations of trend items, the Core 3 contractors copied those into the computer-readable XLIFF format used for the computer-based instruments. This was done both as a quality control process and to reduce the tasks assigned to countries given the short development timeframe for the project.

Once the XLIFF files were created, countries were asked to perform a review by comparing the new computer versions with PDF files of their paper-based items that were supplied by the contractors. These PDF files had been assembled for countries by retrieving their existing paper-based materials and organising them into the 2015 clusters. Countries were asked to document any content errors, which included typographical mistakes or text errors introduced in the process of copying and pasting across formats. Any content issues identified by countries were reviewed by verifiers on the linguistic quality control team and, if approved, the verifiers made the needed change in the computer files. If countries identified any serious layout issues, those were reviewed and, where appropriate, corrected by the Core 2 technical team. As an additional quality control check, the Core 3 contractor also performed layout checks of all items in all languages to identify errors that may have been missed.

Because trend items were selected from previous PISA administrations going back as far as 2003, countries that had not participated in all previous cycles did not have translations for some items. Where this occurred, National Centres were responsible for translating that content in a subsequent step in the development process and these materials were treated as new translations. An additional task for all countries was to provide translations for the recurring directions and prompts. Instructions from the paper booklets, such as “Circle either YES or NO” were revised to “Click on either YES or NO”, and some new directions, such as “Click on the NEXT arrow”, had been specifically developed for the computer-based items. All such recurring directions were identified by the contractors and provided to national teams. National translations of these revised or new directions went through the translation verification process and, once verified, were copied into the computer files by Core 3.

#### **Computer-based new items**

All new science, collaborative problem solving and, where applicable, financial literacy items needed to be translated by national teams following the translation and reconciliation processes defined in the PISA standards (see Chapter 5 for detailed information about this process). Following verification of national translations and the corrections of any remaining errors, countries were asked to sign off on their cognitive materials and those files were then considered locked.



### **Preparing the Field Trial National Student Delivery Systems (SDS)**

The Student Delivery System (or SDS) was a self-contained set of applications for delivery of the PISA 2015 CBA assessments and computer-based student questionnaires. A master version was assembled first for countries to test within their national IT structure. This allowed countries to become familiar with the operation of the SDS and to check the compatibility of the software with computers being used to administer the assessment.<sup>10</sup>

Once all components of national materials were approved and locked, including both the questionnaires and the tests, the national SDS was assembled and tested first by Core 2 (responsible for computer platform development). The SDS was then released to countries for national testing. Countries were asked to check their SDS following a specific testing plan provided by Core 2 and to identify any residual content or layout issues. Where issues were identified those were corrected and a second SDS was released. Once countries signed off on their national SDS, their instruments were released for the field trial.

### **Paper-based trend items**

As previously noted, the mode effects study for the PISA 2015 field trial required all countries to administer the 18 paper-and-pencil forms that included the trend items for reading, mathematics and science. National versions of the paper-based trend clusters were prepared by extracting clusters from existing booklets in the PISA archives and formatting them for the 2015 cycle. To better ensure comparability of the paper-based assessment materials across countries and languages, booklets were centrally created by Core 3 and then reviewed and approved by countries. Those countries who were new to PISA 2015 or who were missing some items from previous cycles needed to translate those materials following the standard translation and verification process. All countries needed to update and translate the common booklet parts, which included the cover, general instructions, formula sheet for mathematics and the acknowledgements page.

For computer-based countries, it was important to ensure comparability across the paper-based and computer-based trend items. Thus, clusters for the paper-based booklets were finalised by the contractors once all computer-based materials were locked. Where errors had been identified in any computer-based versions of trend items, those were also corrected by the contractor in the paper-based files. Once paper-based versions were assembled, they were provided for national review. Any remaining errors identified by countries were corrected and countries were asked to sign off on their materials.

The approved clusters were then assembled into the 18 field trial paper booklets by the contractors in a centralised fashion that ensured comparability of layout. Additionally, two financial literacy booklets were assembled. As a final step, booklets were released to countries so that the sequence of clusters within forms could be confirmed and, once approved, print-ready versions were provided to National Centres.

Paper-based countries followed essentially this same process. They were asked to first check their assembled clusters for errors. Once those had been corrected and their paper booklets assembled, they were asked to check and sign off on the final instruments.

### **Field trial coding**

Coding guides for trend items were compiled by Core 3 based on previous national versions. For computer countries, the coding guides were designed so that a single version could be used for coding both the paper and computer instruments. This meant that both paper and computer item IDs were included and, where question wording differed between the paper and computer formats, both versions were shown. Any items where the paper version was human coded but the computer version was automatically scored were also identified.

The development of the coding guide for new science items was informed by cognitive labs conducted by the University of Luxembourg. The English master version of the new science coding guide was released in draft form prior to the coder training meeting in January 2014. Based on discussions at that meeting, the coding guide was finalised and the updated English version, along with the French source version, was released to countries in March 2014, prior to the beginning of the field trial data collection period.

### **Field trial coder training**

The international field trial coder training was held in January 2014 and focused on all domains and all items. The goals of the training included both having attendees develop an in-depth understanding of the coding process for each item, so they would be prepared to train coders in their countries, and reaching consensus about the coding rules to



better ensure consistency of coding within and between countries and across cycles. Trainers reviewed the layout of the coding guides, general coding principles, common problems and guidelines for applying special codes. Sample student responses were provided and attendees were required to code them. Where there were disagreements about coding for a particular item, those were discussed so that all attendees understood, and would be able to follow, the intent of the coding guides. The feedback provided by the National Centres in the Field Trial Review Questionnaire reflected a high level of satisfaction with the coding training.

### **Field trial coder queries**

As was the case during previous cycles, Core 3 set up a coder query service for the 2015 field trial. Countries were encouraged to send queries to the service so that a common adjudication process was consistently applied to all coder questions about constructed-response items. Queries were reviewed and responses provided by domain-specific teams including item developers and, for trend items, by members of the response team from previous cycles.

In addition to responses to new queries, the queries report included the accumulated responses from previous cycles of PISA. This helped foster consistent coding of trend items across cycles. The report was regularly updated and posted for National Centres on the PISA portal as new queries were received and processed.

### **Field trial outcomes**

The PISA 2015 field trial was designed to yield information about the quantity and quality of data collected. More specifically, the goals of the field trial included collecting and analysing information regarding:

- the quantity of data and the impact, if any, that survey operations had on that data
- the operational characteristics of the computer-delivery platform
- the quality of the items including both those items that were newly developed for computer-based delivery and those that were adapted from earlier cycles
- the use of the data to establish reliable, valid, and comparable scales based on item-response theory (IRT) models both in paper- and computer-based versions.

Overall, the field trial achieved all the stated goals. This information was crucial for the selection and assembly of the main survey instruments and for refining survey procedures where necessary.

The field trial analyses were conducted in batches based on data submission dates. Most of the analyses implemented to evaluate the goals noted above were based on data received from countries by 31 July 2014. That included 53 datasets, with eight from countries implementing only the paper-based assessment and 45 from countries using the computer-based assessment, including trend items administered both in paper and computer. The field trial analyses were amended after receiving additional data, which increased the number of countries to 68 by the end of 2014. Details of the field trial analysis are discussed in Chapter 9.

## **MAIN SURVEY**

The PISA 2015 main survey began in March 2015 with early testing countries and ended by mid-December 2015 with the late testing countries. The majority of countries completed the main survey data collection by May. In preparation for the main survey, countries reviewed items based on their performance in the field trial and were asked to identify any serious errors still in need of correction. The Core 3 contractors worked with countries to resolve any remaining issues and prepare the national instruments for the main survey.

### **National item review following the field trial**

The item feedback process began in July 2014 and concluded in October 2014 and was conducted in two phases. The first phase occurred before countries received their field trial data and the second after receipt of their data. This two-phase process was implemented to allow for the most efficient correction of any remaining errors in item content or layout given the extremely short turn around period between the field trial and main survey.

Phase 1 allowed countries to report any linguistic or layout issues that were noted during the field trial, including errors to the coding guides. All requests were reviewed by Core 3 and assigned to one of two categories: serious errors that would be expected to impact item functioning and therefore were corrected immediately; and comments that would be re-evaluated based on the field trial data. Errors in category one were corrected centrally by the contractors.

Following release of the field trial data, countries received their Phase 2 updated item feedback forms that included flags for any items that had been identified as not fitting the international trend parameters. Flagged items were reviewed by national teams. As was the case in Phase 1, countries were asked to provide comments about these specific items where they could identify serious errors. Requests for corrections were reviewed by Core 3 and, where approved, implemented.

### Item selection

The initial selection of items recommended for the main survey was made by the test development team based on item statistics from the field trial, country comments, coverage of the domain as specified in the framework, item format and the assessment design. In addition, as response timing information was available for the computer-based items, it was possible to use that information to develop proposed main survey clusters with balanced average testing times.

The main survey item selection process for new science was also informed by an independent item review. In March 2014, Pearson, the company responsible for overseeing the development of the PISA 2015 frameworks as the Core 1 contractor, was commissioned by the OECD secretariat to manage an independent review of the 2015 scientific literacy item pool. The purpose of this review was to gather validity evidence of the alignment and accuracy of new and trend science items in relation to the PISA 2015 framework and to ensure that the main survey pool would be a good representation of the construct. The agreement rate between the reviewers and item developers for the metadata coding of the items was 97%. The review concluded that the science items developed for PISA 2015 covered the framework for scientific literacy as it was intended by its developers and approved by the PGB. In addition, the reviewers found that the items were of high quality. Where there were concerns expressed about individual items, those were reviewed by the item development team and expert group.

National Centres were asked to provide feedback about the proposed main survey item pool during Phase 1 of the national item review process. Comments were due prior to the meeting of the science expert group so they could be considered as part of the SEG's review of the item pool.

In October 2014, the SEG met to review and finalise the proposed item pool for the main study. The experts reviewed the tentative selection, along with a pool of potential alternate items. As a result of their discussions, a small number of items were dropped from the recommended pool and replaced by alternate items.

As part of this process, the SEG also approved the recommended set of released items. All items released following the field trial were taken from the pool of potential alternate items. These items performed well enough in the field trial to be considered for inclusion in the main survey but were not used simply because there were many more items available than were needed to meet the various goals for the main survey item pool.

The item counts for science, mathematics, reading, collaborative problem solving, and financial literacy in both the field trial and main survey are presented in Figure 2.15.

■ Figure 2.15 ■

**Item counts (field trial and main survey) by domain and delivery mode**

Domain	Field trial		Main survey	
	Paper-based	Computer-based	Paper-based	Computer-based
Science	91	304 (213 New, 91 Trend)	85	184 (99 New, 85 Trend)
Mathematics	83	82	83	81
Reading	103	103	103	103
CPS	NA	153	NA	117
Financial literacy	NA	43	NA	43

As Figure 2.15 shows, a number of trend items were dropped between the field trial and main survey or not included in the main survey analysis.

- Two mathematics items were not included in the main study data analysis for computer-based countries. One item could only be administered in paper and so was not used on the computer in either the field trial or main survey. One additional item was dropped due to problems with the computer-based scoring.<sup>11</sup>
- Six trend science items were dropped from the computer-based test and not included in the analysis in both modes. Item parameters for two of those items were not available for 2006 when they were last used, so they could not be



used as trend items.<sup>12</sup> One item had been dropped at the international level in 2003 and so should not have been included in 2015.<sup>13</sup> Finally three items, last used in 2003, did not work well in the field trial and so were not moved forward to the main survey.<sup>14</sup>

- Four CPS items were dropped during main survey data analysis.<sup>15</sup> Additionally, a number of items in each unit were combined, based on the main survey analysis and/or to reflect the branching logic within units. That branching meant that, based on the path students took, they might not see all items in a unit and therefore items needed to be clustered in order to function psychometrically.

### **Construct coverage**

The set of items for the main survey was balanced in terms of construct representation, based on the overall distributions recommended in the frameworks.

A total of 184 items was selected for science, with the distribution as shown in Figure 2.16 below.

■ Figure 2.16 ■  
**Science item counts by framework category<sup>16</sup>**

Competency	Items	Percent	Framework goal
Evaluate and design scientific enquiry	39	21%	20-30%
Explain phenomena scientifically	89	48%	40-50%
Interpret data and evidence scientifically	56	31%	30-40%
<b>Knowledge</b>			
Content	98	53%	54-66%
Epistemic	26	14%	10-22%
Procedural	60	33%	19-31%
<b>System</b>			
Earth and Space	49	27%	28%
Living	74	40%	36%
Physical	61	33%	36%

The 117 items selected in the collaborative problem solving domain were distributed among the framework categories as shown below in Figure 2.17.

■ Figure 2.17 ■  
**Collaborative problem solving item counts by framework category**

CPS Competency	Items	Percent	Framework goal
Establishing and maintaining shared understanding	61	52%	40-50%
Taking appropriate action to solve the problem	26	22%	20-30%
Establishing and maintaining team organisation	30	26%	30-35%
<b>Problem solving process</b>			
Exploring and understanding	22	50%	Approx. 40% (combined)
Representing and formulating	37		
Planning and executing	35	30%	Approx. 30%
Monitoring and reflecting	23	20%	Approx. 30%

### **Preparing the main survey national student delivery systems (SDS)**

The process for creating the main survey national student delivery system (SDS) followed that used during the field trial, beginning with assembly and testing of the master SDS followed by the process for assembling national versions of the main survey SDS.

After all components of national materials were locked, including the questionnaires and cognitive instruments, the student delivery system was assembled and tested by Core 2. Countries were then asked to check their SDS and identify any remaining content or layout issues. Once countries signed off on their national SDS, their instruments were released for the field trial.



## Main survey coding

The process used for the main survey coding training was slightly different from that employed prior to the field trial. Full training was provided for all science items, as the major domain. Based on the reliability results from the field trial, a decision was made to conduct a tailored coding training for a selected set of reading items and not to repeat training for trend mathematics and financial literacy items.

The coder query service was again used in the main survey as it had been in the field trial to assist countries in clarifying any uncertainty around the coding process or responses. Queries were reviewed and responses provided by domain-specific teams including item developers and members of the response team from previous cycles.

## Review of main survey item analyses

The main survey data went through extensive analyses implemented through multistep procedures to ensure the quality of the results. The first steps were implemented to evaluate the overall quality of the data submitted by countries looking at how well the assessment design and booklet assignment were reflected in the data as well as looking for the effects of any possible threats to data quality such as technical problems, scoring inconsistencies, issues related to time limits, and other administration problems. These were followed by more specific analyses including item analysis, coding and treatment of missing data, item response theory scaling including international item fit and item-by-country interactions, conditioning models and generation of plausible values. These procedures are described in more detail in Chapters 9, 10 and 12. Finally, the outcomes of these analyses guided decisions around data products and treatment of items as described in detail in Chapter 19.

## Released items

As has been the case in previous PISA cycles, a number of items were released into the public domain at the time of publication of the PISA 2015 results to illustrate the kinds of items included in the assessment. This was particularly important for this cycle due to the shift from paper to computer as the primary mode of assessment. The OECD decided to release four science units from the main survey in their interactive mode: i) *Sustainable Fish Farming* (3 items), ii) *Bird Migration* (3 items), iii) *Slope-Face Investigation* (2 items), and iv) *Meteoroids and Craters* (4 items). In addition, it decided to release one of the field trial units, *Running in Hot Weather* (6 items), to illustrate the interactive simulation units developed for science. These units are available at [www.oecd.org/pisa](http://www.oecd.org/pisa).

## Notes

1. Consistent with previous cycles, easier and harder forms were developed. Clusters R06a and M06a were used to assemble forms for countries selecting the standard forms while clusters R06b and M06b were used to assemble forms for countries selecting the easier forms.
2. Countries chose at the national level whether they wanted to use the easier or standard mathematics and reading clusters.
3. This range was selected to circumvent a requirement of the software used for this selection, and to ensure equal distribution of the different combinations across the sample.
4. For a more detailed description of the science framework, as well as the adaptations made to the frameworks for the 2015 minor domains, please see OECD (2017), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264281820-en>.
5. The orientation of these panes was reversed for right-to-left languages.
6. Note that one item was used only in the paper-based assessment as it required students to draw a line on a graph – something that could not easily be replicated in the computer-based mode. Thus there were 83 trend items in PBA and 82 in CBA for mathematics.
7. A total of six science trend items, four items last used in 2003 items and the two last used in 2006, were dropped following the field trial.



8. The number of field trialed items was particularly large in science to allow for the possibility of an adaptive assessment in the main survey.
9. See Chapter 5 for additional detail about the translatability assessment.
10. More information about the Student Delivery System is provided in Chapter 18.
11. Item DM155Q01C was the paper-based only item and DM192Q01C was dropped from the main survey analysis on computer.
12. Items DS456Q01C and DS456Q02C.
13. Item DS327Q02C.
14. Items DS133Q01C, DS133Q03C and DS133Q04C.
15. The dropped CPS items include: CC104104 and CC104303 in Meeting in the Park, CC102208 in The Field Trip and CC105405 in The Garden.
16. As noted in Chapter 9, the classification of one item (DS648Q05C) was corrected from "Interpret data and evidence scientifically" to "Explain phenomena scientifically" after scaling. The numbers shown here reflect that correction.

## **References**

- Curran, P. J.** et al. (2008), "Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis", *Developmental Psychology*, Vol. 44/2, pp.365-380, <http://doi.org/10.1037/0012-1649.44.2.365>.
- Glas, C. and K. Jehangir** (2013), "Modeling country-specific differential item functioning", in L. Rutkowski, M. von Davier and D. Rutkowske (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, CRC Press, Taylor and Francis Group, Boca Raton, FL.
- Glas, C. A. W. and N.D. Verhelst** (1995), "Testing the Rasch Model", in G. H. Fisher and I. W. Molenaar (eds.), *Rasch models: Foundation, recent developments, and applications*, pp. 69-96, Springer-Verlag, New York.
- Mazzeo, J., and M. von Davier** (2013), "Linking scales in international large-scale assessments", in L. Rutkowski, M. von Davier and D. Rutkowske (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, CRC Press, Taylor and Francis Group, Boca Raton, FL.
- Meredith, W. and J.A. Teresi** (2006), "An essay on measurement and factorial invariance", *Medical Care*, Vol. 44/11, pp. 69-77.
- Meredith, W.** (1993), Measurement invariance, factor analysis, and factorial invariance, *Psychometrika*, Vol. 58/4, pp. 525-543.
- 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, pp. 1-21.
- Oliveri, M. E., and M. von Davier** (2011), "Investigation of model fit and score scale comparability in international assessments", *Psychological Test and Assessment Modeling*, Vol. 53/3, pp. 315-333.
- Reise, S. P., K. F. Widaman and R. H. Pugh** (1993), "Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance", *Psychological Bulletin*, Vol. 114/3, pp. 552-566.





3

# Context questionnaire development

<b>Introduction .....</b>	58
<b>The PISA context questionnaire framework.....</b>	58
<b>The PISA 2015 context questionnaires .....</b>	60
<b>Quality assurance in the development of questionnaires .....</b>	62



## INTRODUCTION

The context questionnaires in PISA provide information on the learning context at the individual, school, and education system or country/economy level. They assess non-cognitive outcomes, individual dispositions and structural and process characteristics of the institutional context. This diverse set of constructs is measured by addressing various stakeholders, namely students and school principals in all countries and economies, as well as parents and teachers in countries and economies that choose to implement additional optional questionnaires.

The questionnaire development for the sixth cycle of PISA introduced several innovations:

- a modular approach for the questionnaire design to identify (a) policy issues which participating countries and economies wanted to be addressed, (b) conceptual constructs related to the respective policy issue, and (c) measures (individual items, indices or questionnaire scales) operationalising these constructs
- an attempt to identify core questionnaire content which needs to be covered across cycles of PISA to report trends in education, finding a balance between core measures and new measures dealing with topics that are important for current education policy
- transitioning the context questionnaires from paper administration to computer-based administration mode
- a teacher questionnaire as an international option.

This chapter provides a brief overview of the questionnaires and their development process, while Chapter 16 describes the questionnaire scaling approaches and index construction and Chapter 17 describes the questionnaire design and implementation into the electronic platform. For more detailed information about different steps of instrument development and how the field trial informed the final instruments see also Kuger et al. (2016).

## THE PISA CONTEXT QUESTIONNAIRE FRAMEWORK

Questionnaire development in PISA has been guided by different approaches since the first questionnaire framework was published for PISA 2009. While previous frameworks focussed on the hierarchical structure of educational systems (PISA 2009) and questions of educational effectiveness (2012), the framework and questionnaire development for PISA 2015 aimed at combining the existing approaches with new aspects of policy interest that currently guide the discussion on educational effectiveness and education policy decisions. Consequently, the questionnaire development used an iterative process linking policy demands with education research foundations and possibilities for instrument implementation.

The starting point for development of the PISA 2015 questionnaire framework (OECD, 2017) was a proposal for 19 highly important policy issues (so-called modules). These modules included aspects of science education, equity, broader educational outcomes beyond achievement, supportive school context and educational governance. The modules are presented in Figure 3.1. As a first step, each module was defined and explored based on literature from educational research and experience in previous cycles of PISA. The members of the PISA Governing Board (PGB) were then asked to provide feedback on the modules' definitions and rate their importance for reporting.

The areas which received the highest policy relevance included non-cognitive outcomes (modules 4 and 10), teaching and learning (modules 1, 2, and 12), and school policies (modules 15 and 19). This indication of policy relevance formed the basis of the development of questionnaire material, i.e. based on these ratings, trend material repeated from previous cycles was integrated and new material was developed for high-priority modules allowing more in-depth assessment in the field trial in PISA 2015 (see Chapter 17 for the design).

Another underlying principle in instrument development was balancing trend and new reporting on additional aspects of learning contexts. As one of the aims of PISA is to measure trend indicators across cycles, the framework identified the core content of questionnaire material, i.e. constructs of context assessment that should be kept across all cycles. This material was granted higher priority in instrument development to enable later trend reporting. All of the core content as displayed in Figure 3.2 is covered by the PISA 2015 questionnaires, mostly taking up measures from previous cycles, especially – for science-related constructs – from PISA 2006.

For PISA 2015, the conceptual framework for the context questionnaires has already been published (OECD, 2017). Therefore, this chapter only provides a summary of the context questionnaire framework and the questionnaire development. The newly-developed material that was not taken over into the main survey, but only used in the field trial, has been documented by Kuger et al. (2016).



■ Figure 3.1 ■

### Modular structure of the PISA 2015 questionnaire design

Student background		Processes			Non-cognitive outcomes
Family	Education	Actors	Core processes	Resource allocation	
Science-related topics	5. Out-of-school science experience	1. Teacher qualification and professional knowledge	2. Science teaching practices	12. Learning time and curriculum	4. Science-related outcomes: motivation, interest, beliefs...
		<b>Teaching and learning</b> 3. School-level learning environment for science			
General topics	7. Student SES and family  8. Ethnicity and migration	14. Parental involvement  15. Leadership and school management	13. School climate: interpersonal relations, trust, expectations	16. Resources	6. Career aspirations  10. General behaviour and attitudes  11. Dispositions for collaborative problem solving
	9. Educational pathways in early childhood	17. Locus of decision making within the school system	19. Assessment, evaluation and accountability	18. Allocation, selection and choice	
		<b>School policies</b> <b>Governance</b>			

Source: OECD (2017), "Modular structure of the PISA 2015 context assessment design", in *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*.

■ Figure 3.2 ■

### Constructs identified as core content in the PISA 2015 Questionnaire Framework

	Student and school background	Processes	Non-cognitive outcomes
System Level		<b>Governance:</b> Decision making, horizontal and vertical differentiation	(Aggregated student data)
School Level	School location Type and size of school Amount and source of resources (incl. ICT) Social/ethnic/academic composition Class size Teacher qualification	<b>School policies:</b> Programmes offered, admission and grouping policies Allocated and additional learning time <i>Extra-curricular activities</i> , Professional development, leadership, parental involvement Assessment/evaluation/accountability policies School climate (teacher and student behaviour) <b>Teaching and learning:</b> Disciplinary climate, teacher support, cognitive challenge	(Aggregated student data) Drop-out rate
Student Level	Gender SES Language and migration background Grade level Pre-primary education Age at school entry	Grade repetition Programme attended Learning time at school (mandatory lessons and additional instruction) <i>Out-of school learning</i>	Domain-general non-cognitive outcomes (e.g. achievement motivation, well-being in school) Domain-specific non-cognitive outcomes ( <i>motivation, domain-related beliefs and strategies, self-related beliefs, domain-related behaviour</i> )

Note: Measures in italics were adapted to the major domain, i.e. science in PISA 2015.

Source: OECD (2017), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*.

As in previous cycles, the Questionnaire Expert Group (QEG) guided the development of the PISA context questionnaires and framework through regular meetings. The members reviewed questionnaire drafts as well as feedback from countries and economies and discussed the material together with the OECD Secretariat and the international contractors to ensure the link between the assessment, the context questionnaires, and the frameworks. For the QEG 2015, liaison persons were nominated to attend meetings of the Science Expert Group and the Expert Group for Collaborative Problem Solving. This guaranteed a close link between the development of the assessment framework and tests and the questionnaire development process.

## THE PISA 2015 CONTEXT QUESTIONNAIRES

The following questionnaires were administered in the PISA 2015 main survey:

- the Student Questionnaire (computer-based and paper-based)
- the School Questionnaire (computer-based and paper-based)
- the Educational Career Questionnaire as an international option (computer-based)
- the ICT Familiarity Questionnaire as an international option (computer-based)
- the Parent Questionnaire as an international option (paper-based)
- the Teacher Questionnaire as an international option (computer-based).

One important guiding principle for the development of the PISA 2015 questionnaires was that all policy modules (see Figure 3.1) should be represented in several questionnaires, thus gathering important information from different, and if possible the most knowledgeable, sources. Field trial data were used to choose the most reliable approach and source of information for each construct and module. Figure 3.3 highlights the coverage of policy issues across questionnaires for the final main survey questionnaires.

■ Figure 3.3 ■

### Overview of the 19 policy issues (modules) and their relation to the questionnaires

		STQ	SCQ	TCQ	PAQ	ICT	EC
<b>Policy area: Science education</b>							
1.	Teacher qualification and professional knowledge		X	X			
2.	Science teaching practices	X	X	X			
3.	School-level learning environments for science		X	X		X	X
4.	Science-related outcomes: motivation, attitudes, beliefs, strategies	X				X	
5.	Out-of-school science experience	X			X	X	X
6.	Career aspirations	X					
<b>Policy area: Equity</b>							
7.	Student SES, family and home background	X			X	X	
8.	Ethnicity and migration	X		X	X		
9.	Educational pathways in early childhood	X			X		X
<b>Policy area: Broader educational outcomes beyond achievement</b>							
10.	Domain-general student behaviour and attitudes	X				X	
11.	Student dispositions related to collaborative problem solving	X		X			
<b>Policy area: Supportive school context</b>							
12.	Learning time and curriculum	X	X	X			X
13.	School climate: Interpersonal relations, trust, expectations	X	X				
14.	Parental involvement	X	X		X		
15.	Leadership and school management		X	X			
16.	Resources		X	X		X	
<b>Policy area: Educational governance</b>							
17.	Locus of control within the school system		X				
18.	Allocation, selection and choice		X		X		
19.	Assessment, evaluation and accountability	X	X	X			

Note: The following acronyms are used for: Student Questionnaire (STQ), School Questionnaire (SCQ), Teacher Questionnaire (TCQ), Parent Questionnaire (PAQ), ICT Familiarity (ICT) and Educational Career (EC). X indicates if this module was implemented in the respective instrument.

### The Student Questionnaire (computer-based and paper-based)

As in previous cycles, the PISA Student Questionnaire was administered to all students participating in the PISA assessment. It was administered on computer, while countries testing on paper implemented a slightly shorter version.

### The School Questionnaire (computer-based and paper-based)

As in previous cycles, the PISA School Questionnaire was administered to the principal for those schools participating in PISA. It was administered on computer, while countries and economies using paper-based testing implemented a slightly shorter version.



## **The Educational Career Questionnaire (computer-based)**

This optional questionnaire was first introduced in 2003 and was administered to all students participating in PISA if a country or economy chose to implement this option. It included additional questions on students' past and current education, focussing on additional instruction and learning time in PISA 2015. The Educational Career option was administered after the main Student Questionnaire.

## **The ICT Familiarity Questionnaire (computer-based)**

This optional questionnaire was first introduced in PISA 2003 and was administered to all students participating in PISA if a country or economy chose to implement this option. It included additional questions on students' usage of electronic and digital devices, as well as their confidence and attitudes towards ICT. The ICT option was administered after the main Student Questionnaire.

## **The Parent Questionnaire (paper-based)**

The optional Parent Questionnaire was administered on paper and targeted the parents of all students participating in PISA. It enquired about learning contexts, support, and resources at home as well as spending on education and parents' science-related interests and attitudes.

## **The Teacher Questionnaire**

The Teacher Questionnaire was introduced for the first time in PISA 2015. The underlying idea was that important predictors of academic achievement, such as teacher qualification and quality of teaching and learning settings, are best assessed by asking teachers directly. Resulting data can be used to analyse differences between countries/economies and schools. Although some of these aspects were also covered by the School Questionnaire or the Student Questionnaire, administering a questionnaire to teachers was likely to improve the objectivity, reliability, and validity of information. Teachers were addressed as experts for teaching and student learning in the Teacher Questionnaire. The framework and item development for the Teacher Questionnaire were integrated into the overall development process of the PISA questionnaires, thus fitting in with the overall design and the policy issues mentioned above.

Taking into account the major domain of science as well as general differences in teacher characteristics and practices, PISA 2015 implemented two different teacher questionnaires. One questionnaire addressed teachers eligible for teaching science to 15-year-olds in PISA schools, the other one addressed teachers of all other subjects. For detailed information about the sampling see Chapter 4.

Implementing a Teacher Questionnaire into PISA yields several opportunities, as it can deliver information on:

- the professional background of teachers
- the education and training of teachers, including school-based professional development
- teachers' beliefs and attitudes
- school level policies such as teacher co-operation, and shared values
- teachers' perception of school culture, school management and leadership, parental involvement, and school development
- domain-specific and domain-general instructional policies and practices
- the curriculum and opportunity-to-learn.

The PISA 2015 Teacher Questionnaire focussed on the policy topics described below.

### **Teacher qualification and professional knowledge (module 1)**

While basic information on teacher qualification is available from the School Questionnaire, the Teacher Questionnaire incorporated questions that were partially taken from the OECD Teaching and Learning International Study (TALIS) (OECD, 2009). This includes teacher background information, such as gender, age, employment status, job experience, information on initial education and professional development, as well as information about teachers' beliefs, self-efficacy (for example on teaching science), and their job satisfaction.



### **Science teaching practices (module 2) and school-level learning environments for science (module 3)**

Science teachers were asked to describe their teaching practices in two longer questions: “*Teacher-directed teaching and learning activities in science lessons*” and a selected set of “*Inquiry-based activities*”. As parallel scales were implemented in the Student Questionnaire, teacher and student perspectives could be combined and compared (triangulated) at school level. In addition, teachers reported about collaborative learning as well as assessment and grading practices in the classroom.

### **Learning time and curriculum (module 12)**

Teachers are the stakeholders who can be assumed to be most knowledgeable of the science curriculum. They were thus asked whether there was a formal curriculum in place, which educational goals and processes were covered in the curriculum and whether the students’ parents were informed about the curriculum.

### **Leadership and school management (module 15)**

The Teacher Questionnaire also collected information on school leadership and management from teachers’ perspectives. These questions covered aspects of the principal’s leadership style.

### **School resources (module 16) and assessment, evaluation and accountability (module 19)**

Teachers answered a question that reported their perspective on teaching resources in the school and the extent to which they might affect their capacity to provide instruction. This question complements a parallel question in the School Questionnaire. In addition, teachers were asked about their experiences with school evaluation.

The Teacher Questionnaire was implemented as an international option and was administered via an electronic online platform. Teachers were given individual access to this platform, providing each eligible teacher within a school with an individual password. This procedure guaranteed nondisclosure of teacher identity to any stakeholder, including the school principal. Chapter 17 explains the technical implementation in more detail.

## **QUALITY ASSURANCE IN THE DEVELOPMENT OF QUESTIONNAIRES**

Specific standards underlie the PISA questionnaire development process as well as the implementation of the material into the final instruments. These standards aim at quality assurance as well as comparability of the data across countries and economies. Mechanisms for PISA 2015 included a national review, cognitive labs, linguistic translatability assessment, centralized transfer of trend material, negotiation of adaptations and linguistic verification. The following sections each give a short introduction to these procedures.

### **National review**

PISA questionnaires aim at covering topics of education that are important to all participating countries and economies and that can help to explain student achievement both within and between countries/economies. To achieve this goal, newly developed material was shared with representatives of countries and economies at an early stage in the development process to obtain in-depth feedback. This process not only helps to ensure comparability, but asks for ratings on several important factors for each question to be implemented in PISA. Each participating country and economy was asked to judge the relevance of the specific topic for their educational system. The review also aimed to establish whether the addressee that is targeted in the questionnaire (e.g. teachers, principals) is indeed the best person to answer. A very important aspect of ratings touched on issues of sensitivity. Feedback was collected on whether a topic might be sensitive, i.e. was politically acceptable, complied with data privacy regulations in the country/economy or could lead to cultural bias. Potential translation and adaptation difficulties were also addressed in this review. Finally, countries and economies were asked to give an overall rating of each proposed question. Based on these national reviews, proposed questions were rephrased or even deleted.

### **Cognitive labs**

Newly developed questionnaire material for the Student and School questionnaire was pre-tested in English and French, and in English, French and Spanish for the Teacher Questionnaire during the development stage. This pre-testing was implemented in the form of cognitive labs with small groups of students and teachers. The respondents first answered selected, newly-developed questions. During this phase, the test administrator recorded the time it took to read and answer the questions. In a second step, respondents were asked about the answering process including whether they understood the questions, if they could answer these based on the response options given and about any other comment



they might want to give. In addition, small focus group interviews were conducted with teachers to discuss the newly-developed Teacher Questionnaire material. All feedback was collected and led to revision of the proposed questionnaire material.

### **Translatability assessment**

To enhance comparability, a translatability assessment of the questionnaire material was carried out. Linguistic experts evaluated the material with due consideration for the Ask-the-Same-Question (ASQ) model (Harkness, 2003). This approach seeks to optimize the wording in the source questionnaire so that the items can be translated in all relevant languages while maintaining the construct covered, and therefore maintaining the intended measurement properties. The newly-developed questionnaire material was translated into several languages representing the most common language groups, including an East-Asian language (Korean), a Slavic language, an Indo-German language (German), a Romance language (French), and Modern Standard Arabic. Translators highlighted any linguistic issues related to the translation of the questionnaire content that could lead to non-translatability or possible bias in later meaning of a question. Questionnaire developers then revised the material based on this feedback. The translatability assessment is described in detail in Chapter 5 of this report.

### **Centralised trend material transfer**

With the transition to computer-based assessment, the international contractors implemented a centralized transfer process for national trend material. All questionnaire material from previous cycles that was chosen to be administered again for PISA 2015 was centrally transferred into the electronic platform by Core 3. Any changes to these questions needed to be requested and justified by the country/economy. This process allowed for external control to preserve national trend material in PISA 2015. For more explanation see Chapter 5 and Chapter 17.

### **Adaptation negotiation and verification**

In some cases, cultural traditions, national understanding of a question or features of the education system vary largely, leading to the need for adaptations in questionnaires. As in previous cycles, the National Centres in each country and economy were asked to document which national adaptations they needed or wished to implement in the materials by describing them in specially designed standardized forms. For the questionnaires, a Questionnaire Adaptation Spreadsheet (QAS) was provided describing all adaptations that a country or economy wished to implement. For each country/economy and each questionnaire, all adaptations were checked by the international contractors and documented in the QAS. After translation and negotiation of adaptations, all national material was verified by the international contractors. Linguistic checks were performed, and any unclear translation was discussed with the international questionnaire developers, the country/economy, and the linguistic quality control team (Core 3). More information is given in Chapter 5.

All final questionnaire material was then implemented into the paper-based or computer-based versions, tested, and provided to the PISA participants. Further information about these steps is given in Chapter 17.

### **References**

Harkness, J. A. (2003), Questionnaire Translation, in Harkness, J. A. et al. (Eds.), *Cross-Cultural Survey Methods*, Wiley, Hoboken.

Kugler, S. et al. (eds.) (2016), *Assessing Contexts of Learning World-Wide*, Springer, Berlin.

OECD (2017), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264281820-en>.

OECD (2009), *Creating Effective Teaching and Learning Environments: First results From TALIS*, OECD Publishing, Paris.





## 4

# Sample design

<b>Target population and overview of the sampling design .....</b>	66
<b>Population coverage, and school and student participation rate standards.....</b>	67
<b>Main study school sample.....</b>	70
<b>Student samples .....</b>	84
<b>Teacher samples .....</b>	86
<b>Definition of school.....</b>	86

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

## TARGET POPULATION AND OVERVIEW OF THE SAMPLING DESIGN

The desired base PISA target population in each country consisted of 15-year-old students attending educational institutions in grades 7 and higher. This meant that countries were to include:

- 15 year olds enrolled full-time in educational institutions
- 15 year olds enrolled in educational institutions who attended only on a part-time basis
- students in vocational training programmes, or any other related type of educational programmes
- students attending foreign schools within the country (as well as students from other countries attending any of the programmes in the first three categories).

It was recognised that no testing of 15 year olds schooled in the home, workplace or out of the country would occur and therefore these 15 year olds were not included in the international target population.

The operational definition of an age population directly depends on the testing dates. The international requirement was that the assessment had to be conducted during a 42-day period, referred to as the testing period, between 1 March 2015 and 31 August 2015, unless otherwise agreed.

Further, testing was not permitted during the first six weeks of the school year because of a concern that student performance levels may have been lower at the beginning of the academic year than at the end of the previous academic year, even after controlling for age.

The 15-year-old international target population was slightly adapted to better fit the age structure of most Northern Hemisphere countries. As the majority of the testing was planned to occur in April, the international target population was consequently defined as all students aged from 15 years and 3 completed months to 16 years and 2 completed months at the beginning of the assessment period. This meant that in all countries testing in April 2015, the target population could have been defined as all students born in 1999 who were attending an educational institution, as defined above.

A variation of up to one month in this age definition was permitted. This allowed a country testing in March or in May to still define the national target population as all students born in 1999. If the testing took place between June and December, the birth date definition had to be adjusted so that in all countries the target population always included students aged 15 years and 3 completed months to 16 years and 2 completed months at the time of testing, or a one month variation of this.

In all but one country, the Russian Federation, the sampling design used for the PISA assessment was a two-stage stratified sample design. The first-stage sampling units consisted of individual schools having 15-year-old students, or the possibility of having such students at the time of assessment. Schools were sampled systematically from a comprehensive national list of all PISA-eligible schools, known as the school sampling frame, with probabilities that were proportional to a measure of size. The measure of size was a function of the estimated number of PISA-eligible 15-year-old students enrolled in the school. This is referred to as systematic probability proportional to size (PPS) sampling. Prior to sampling, schools in the sampling frame were assigned to mutually exclusive groups based on school characteristics called explicit strata, formed to improve the precision of sample-based estimates.

The second-stage sampling units in countries using the two-stage design were students within sampled schools. Once schools were selected to be in the sample, a complete list of each sampled school's 15-year-old students was prepared. Each country had to set a target cluster size (TCS) of 42 students for computer-based countries and 35 for paper-based countries, although with agreement countries could use alternative values. The sample size within schools is prescribed, within limits, in the PISA Technical Standards (see Annex F). From each list of students that contained more than the target cluster size, a sample of around 42 students were selected with equal probability and for lists with fewer than the target number, all students on the list were selected.

The target cluster size remained the same for countries participating in the international option of financial literacy (FL) in 2015, as the students selected for this assessment were a subsample of the students sampled for the regular PISA test (see Chapter 2).

In the Russian Federation, a three-stage design was used. In this case, geographical areas were sampled first (first-stage units) using probability proportional to size sampling, and then schools (second-stage units) were selected within these



sampled geographical areas. Students were the third-stage sampling units in this three-stage design and were sampled from the selected schools.

## **POPULATION COVERAGE, AND SCHOOL AND STUDENT PARTICIPATION RATE STANDARDS**

To provide valid estimates of student achievement, the sample of students had to be selected using established and professionally recognised principles of scientific sampling in a way that ensured representation of the full target population of 15-year-old students in the participating countries.

Furthermore, quality standards had to be maintained with respect to (i) the coverage of the PISA international target population, (ii) accuracy and precision, and (iii) the school and student response rates.

### **Coverage of the PISA international target population**

National Project Managers (NPMs) might have found it necessary to reduce their coverage of the target population by excluding, for instance, a small, remote geographical region due to inaccessibility, or a language group, possibly due to political, organisational or operational reasons, or special education needs students. Areas deemed to be part of a country (for the purpose of PISA), but which were not included for sampling, although this occurred infrequently, were designated as non-covered areas. Care was taken in this regard because, when such situations did occur, the national desired target population differed from the international desired target population. In an international survey in education, the types of exclusion must be defined consistently for all participating countries and the exclusion rates have to be limited. Indeed, if a significant proportion of students were excluded, this would mean that survey results would not be representative of the entire national school system. Thus, efforts were made to ensure that exclusions, if they were necessary, were minimised according to the PISA 2015 Technical Standards (see Appendix F).

Exclusion can also take place either at the school level (exclusion of entire schools) or at the within-school level (exclusion of individual students) often for special education needs or language. International within-school exclusion rules for students were specified as follows:

- Intellectually disabled students are students who have a mental or emotional disability and who, in the professional opinion of qualified staff, are cognitively delayed such that they cannot be validly assessed in the PISA testing setting. This category includes students who are emotionally or mentally unable to follow even the general instructions of the test. Students could not be excluded solely because of poor academic performance or normal discipline problems.
- Functionally disabled students are students who are permanently physically disabled in such a way that they cannot be validly assessed in the PISA testing setting. However, functionally disabled students who could provide responses were to be included in the testing.
- Students with insufficient assessment language experience are students who need to meet all of the following criteria: i) are not native speakers of the assessment language(s), ii) have limited proficiency in the assessment language(s), and iii) have received less than one year of instruction in the assessment language(s). Students with insufficient assessment language experience could be excluded.
- Students not assessable for other reasons as agreed upon. A nationally-defined within-school exclusion category was permitted if agreed upon by the international contractor. A specific subgroup of students (for example students with severe dyslexia, dysgraphia, or dyscalculia) could be identified for whom exclusion was necessary but for whom the previous three within-school exclusion categories did not explicitly apply, so that a more specific within-school exclusion definition was needed.
- Students taught in a language of instruction for the main domain for which no materials were available. Standard 2.1 notes that the PISA test is administered to a student in a language of instruction provided by the sampled school in the major domain of the test. Thus, if no test materials were available in the language in which the sampled student is taught, the student was excluded. For example, if a country has testing materials in languages X, Y, and Z, but a sampled student is taught in language A, then the student can be excluded since there are no testing materials available in the student's language of instruction.

A school attended only by students who would be excluded from taking the assessment for intellectual, functional, or linguistic reasons was considered a school-level exclusion.

The overall exclusion rate within a country (i.e. school-level and within-school exclusions combined) needed to be kept below 5% of the PISA desired target population. Guidelines for restrictions on the level of exclusions of various types were as follows:

- School-level exclusions for inaccessibility, feasibility or other reasons were to cover less than 0.5% of the total number of students in the PISA desired target population for participating countries. Schools in the school sampling frame which had only one or two PISA-eligible students were not allowed to be excluded from the frame. However, if, based on the frame, it was clear that the percentage of students in these small schools would not cause a breach of the 0.5% allowable limit, then such schools could be excluded in the field at that time of the assessment, if they still only had one or two PISA-eligible students.
- School-level exclusions for intellectually or functionally disabled students, or students with insufficient assessment language experience, were to cover fewer than 2% of the PISA desired target population of students.
- Within-school exclusions for intellectually disabled or functionally disabled students, or students with insufficient assessment language experience, or students nationally-defined and agreed upon for exclusion were expected to cover less than 2.5% of PISA students. Initially, this could only be an estimate. If the actual percentage was ultimately greater than 2.5%, the exclusion percentage was re-calculated without considering students who were excluded because of insufficient familiarity with the assessment language as this is a largely unpredictable part of each country's PISA-eligible population, not under the control of the education system. If the resulting percentage was below 2.5%, the exclusions were regarded as acceptable. Otherwise the level of exclusion was given consideration during the data adjudication process, to determine whether there was any need to note the results, or take other action in relation to reporting the data.

### **Accuracy and precision**

A minimum of 150 schools were selected in each country; if a participating country had fewer than 150 schools then all schools participated. Within each participating school, a predetermined number of students – the target cluster size (usually 42 students in computer-based countries and 35 students in paper-based countries) – were randomly selected with equal probability. In schools with fewer than number of target cluster size-eligible students, all students were selected. In total, a minimum sample size of 5 250 assessed students was needed in computer-based countries (and 4 500 assessed students in paper-based countries), or the entire population if it was less than this size. It was possible to negotiate a target cluster size that differed from 42 students, but if it was reduced then the sample size of schools was increased to more than 150, so as to ensure that at least the minimum sample size of assessed students would be reached. The target cluster size selected per school had to be at least 20 students, so as to ensure adequate accuracy in estimating variance components within and between schools – a major analytical objective of PISA.

NPMs were strongly encouraged to identify available variables to use for defining the explicit and implicit strata for schools to reduce the sampling variance. See the section “Stratification”, further on in this chapter for more details.

For countries participating in PISA 2012 that had larger than anticipated sampling variances associated with their estimates, recommendations were made regarding sample design changes that would possibly help to reduce the sampling variances for PISA 2015. These included modifications to stratification variables and increases in the required school sample size.

### **School response rates**

A response rate of 85% was required for initially-selected schools. If the initial school response rate fell between 65% and 85%, an acceptable school response rate could still be reached through the use of replacement schools. Figure 4.1 provides a summary of the international requirements for school response rates. To compensate for a sampled school that did not participate, where possible, two potential replacement schools were identified. The school replacement process is described in the section further on in this chapter “School sample selection”.

Furthermore, a school with a student participation rate between 25% and 50% was not considered as a participating school for the purposes of calculating and documenting response rates.<sup>1</sup> However, data from such schools were included in the database and contributed to the estimates included in the initial PISA international report. Data from schools with a student participation rate of less than 25% were not included in the database, and such schools were regarded as non-respondents.

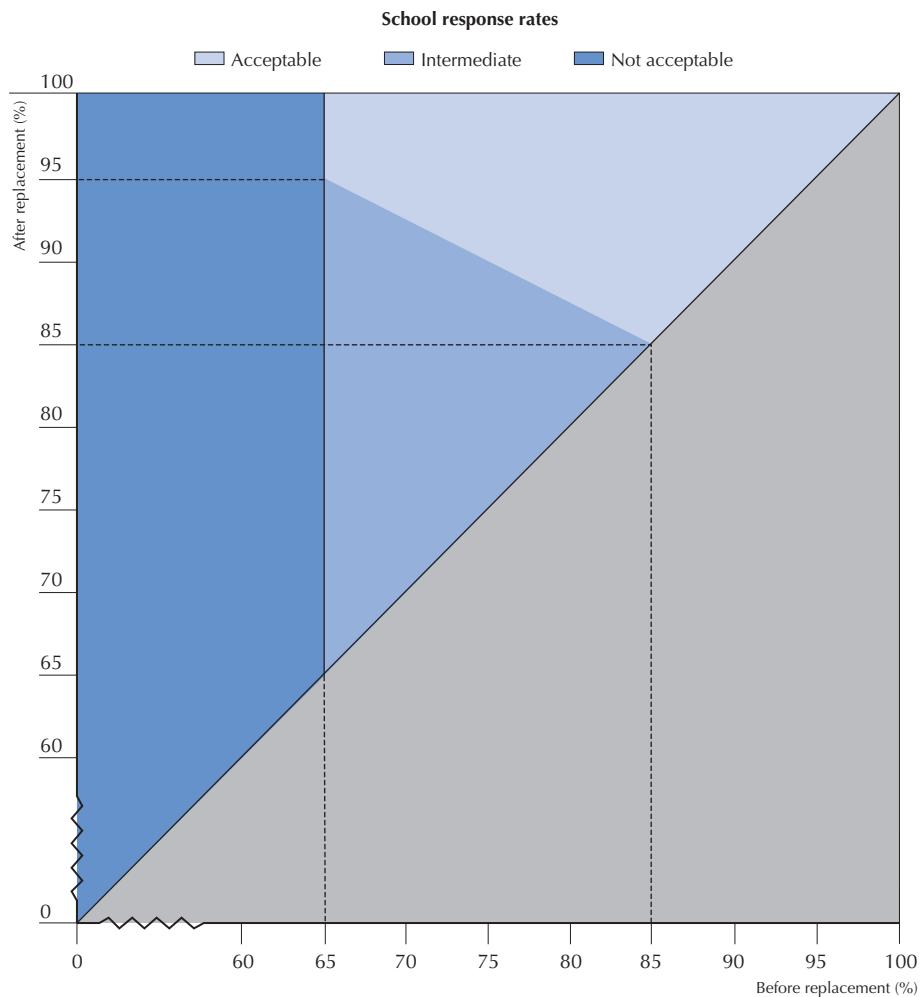
The rationale for this approach was as follows. There was concern that, in an effort to meet the requirements for school response rates, a National Centre might allow schools to participate that would not make a concerted effort to ensure



that students attended the assessment sessions. To avoid this, a standard for student participation was required for each individual school in order that the school be regarded as a participant. This standard was set at a minimum of 50% student participation. However, there were a few schools in many countries that conducted the assessment without meeting that standard. Thus it had to be decided if the data from students in such schools should be used in the analyses, given that the students had already been assessed. If the students from such schools were retained, non-response bias would possibly be introduced to the extent that the students who were absent could have achieved different results from those who attended the testing session, and such a bias is magnified by the relative sizes of these two groups. If one chose to delete all assessment data from such schools, then non-response bias would be introduced as the schools were different from others in the sample, and sampling variance would be increased because of sample size attrition.

It was decided that, for a school with between 25% and 50% student response, the latter source of bias and variance was likely to introduce more error into the study estimates than the former, but with the converse judgement for those schools with a student response rate below 25%. Clearly the cut-off of 25% is arbitrary as one would need extensive studies to try to establish this cut-off empirically. However, it is clear that, as the student response rate decreases within a school, the possibility of bias from using the assessed students in that school will increase, while the loss in sample size from dropping all of the students in the school will be small.

■ Figure 4.1 ■  
**School response rate standards**



These PISA standards applied to weighted school response rates. The procedures for calculating weighted response rates are presented in Chapter 8. Weighted response rates weigh each school by the number of students in the population that are represented by the students sampled from within that school. The weight consists primarily of the enrolment size

of 15-year-old students in the school, divided by the selection probability of the school. Because the school samples were selected with probability proportional to size, in most countries most schools contributed approximately equal weights. As a consequence, the weighted and unweighted school response rates were similar. Exceptions could occur in countries that had explicit strata that were sampled at very different rates. Details as to how each participating economy and adjudicated region performed relative to these school response rate standards are included in Chapters 11 and 14.

### **Student response rates**

An overall response rate of 80% of selected students in participating schools was required. A student who had participated in the original or follow-up cognitive sessions was considered to be a participant. A minimum student response rate of 50% within each school was required for a school to be regarded as participating: the overall student response rate was computed using only students from schools with at least a 50% student response rate. Again, weighted student response rates were used for assessing this standard. Each student was weighted by the reciprocal of his/her sample selection probability.

## **MAIN STUDY SCHOOL SAMPLE**

### **Definition of the national target population**

NPMs were first required to confirm their dates of testing and age definition with the international contractor. Once these were approved, NPMs were notified to avoid having any possible drift in the assessment period leading to an unapproved definition of the national target population.

Every NPM was required to define and describe their country's target population and explain how and why it might deviate from the international target population. Any hardships in accomplishing complete coverage were specified, discussed and approved or not, in advance. Where the national target population deviated from full coverage of all PISA-eligible students, the deviations were described and enrolment data provided to measure how much coverage was reduced. The population, after all exclusions, corresponded to the population of students recorded on each country's school sampling frame. Exclusions were often proposed for practical reasons such as increased survey costs or complexity in the sample design and/or difficult testing conditions. These difficulties were mainly addressed by modifying the sample design to reduce the number of such schools selected rather than to exclude them (see Chapter 8 for further details on weighting). Schools with students that would all be excluded through the within-school exclusion categories could be excluded up to a maximum of 2% of the target population as previously noted. Otherwise, countries were instructed to include the schools but to administer the PISA UH booklet, consisting of a subset of the PISA assessment items, deemed more suitable for students with special needs (see Chapter 2 for further details of the UH booklet). Eleven countries used the UH booklet for PISA 2015.

Within participating schools, all PISA-eligible students (i.e. born within the defined time period and in grades 7 or higher) were to be listed. From this, either a sample of target cluster size students was randomly selected or all students were selected if there were fewer than the number of target cluster size-eligible students (as described in the "Student Sampling" section). The lists had to include students deemed as meeting any of the categories for exclusion, and a variable maintained to briefly describe the reason for exclusion. This made it possible to estimate the size of the within-school exclusions from the sample data.

It was understood that the exact extent of within-school exclusions would not be known until the within-school sampling data were returned from participating schools and sampling weights computed. Participating country projections for within-school exclusions provided before school sampling were known to be estimates.

NPMs were made aware of the distinction between within-school exclusions and non-response. Students who could not take the PISA achievement tests because of a permanent condition were to be excluded and those with a temporary impairment at the time of testing, such as a broken arm, were treated as non-respondents along with other absent sampled students.

Exclusions by country are documented in Chapter 11.

### **The sampling frame**

All NPMs were required to construct a school sampling frame to correspond to their national defined target population. The school sampling frame as defined by the *School Sampling Preparation Manual* would provide complete coverage of



the national defined target population without being contaminated by incorrect or duplicate entries or entries referring to elements that were not part of the defined target population. It was expected that the school sampling frame would include any school that could have 15-year-old students, even those schools which might later be excluded or deemed ineligible because they had no PISA-eligible students at the time of data collection. The quality of the sampling frame directly affects the survey results through the schools' probabilities of selection and therefore their weights and the final survey estimates. NPMs were therefore advised to be diligent and thorough in constructing their school sampling frames.

All but one country used school-level sampling frames as their first stage of sample selection. The *School Sampling Preparation Manual* indicated that the quality of sampling frames for both two and three-stage designs would largely depend on the accuracy of the approximate enrolment of 15 year olds available (ENR) for each first-stage sampling unit. A suitable ENR value was a critical component of the sampling frames since selection probabilities were based on it for both two- and three-stage designs. The best ENR for PISA was the number of currently enrolled 15-year-old students. Current enrolment data, however, were rarely available at the time of school sampling, which meant using alternatives. Most countries used the first-listed available option from the following list of alternatives:

- student enrolment in the target age category (15 year olds) from the most recent year of data available
- if 15 year olds tend to be enrolled in two or more grades, and the proportions of students who are aged 15 in each grade are approximately known, the 15-year-old enrolment can be estimated by applying these proportions to the corresponding grade-level enrolments
- the grade enrolment of the modal grade for 15 year olds
- total student enrolment, divided by the number of grades in the school.

The *School Sampling Preparation Manual* noted that if reasonable estimates of ENR did not exist or if the available enrolment data were out of date, schools might have to be selected with equal probabilities which might require an increased school sample size. However, no countries needed to use this option.

Besides ENR values, NPMs were instructed that each school entry on the frame should include at minimum:

- school identification information, such as a unique numerical national identification, and contact information such as name, address and phone number
- coded information about the school, such as region of country, school type and extent of urbanisation, which would be used as stratification variables.

As noted, a three-stage design and an area-level (geographic) sampling frame could be used where a comprehensive national list of schools was not available and could not be constructed without undue burden, or where the procedures for administering the test required that the schools be selected in geographic clusters. As a consequence, the area-level sampling frame introduced an additional stage of frame creation and sampling (first stage) before actually sampling schools (second stage, with the third stage being students). Although generalities about three-stage sampling and using an area-level sampling frame were outlined in the *School Sampling Preparation Manual* (for example, that there should be at least 80 first-stage units and at least 40 needed to be sampled), NPMs were also informed that the more detailed procedures outlined there for the general two-stage design could easily be adapted to the three-stage design. The only country that used a three-stage design was the Russian Federation, where a national list of schools was not available. The use of the three-stage design allowed for school lists to be obtained only for those areas selected in stage one rather than for the entire country. The NPM for the Russian Federation received additional support with their area-level sampling frame.

## **Stratification**

Prior to sampling, schools were to be ordered, or stratified, in the sampling frame. Stratification consists of classifying schools into similar groups according to selected variables referred to as stratification variables. Stratification in PISA was used to:

- improve the efficiency of the sample design, thereby making the survey estimates more reliable
- apply different sample designs, such as disproportionate sample allocations, to specific groups of schools in states, provinces, or other regions
- ensure all parts of a population were included in the sample
- ensure adequate representation of specific groups of the target population in the sample.

There were two types of stratification used: explicit and implicit. Explicit stratification consists of grouping schools into strata that will be treated independently, as if they were separate school sampling frames. Examples of explicit stratification variables could be states or regions within a country. Implicit stratification consists essentially of sorting the schools uniquely within each explicit stratum by a set of designated implicit stratification variables. Examples of implicit stratification variables could be type of school, urbanisation, or minority composition. Implicit stratification is a way of ensuring a strictly-proportional sample allocation of schools across all the groups used for implicit stratification. It can also lead to improved reliability of survey estimates, provided that the implicit stratification variables being considered are correlated with PISA achievement at the school level (Jaeger, 1984). Guidelines on choosing stratification variables that would possibly improve the sampling were provided in the *FT Sampling Guidelines Manual* (OECD, 2013).

Table 4.1 provides the explicit stratification variables used by each country, as well as the number of explicit strata found within each country. For example, Australia had eight explicit strata using states/territories which were then further delineated by three school types (known as sectors) and also had one explicit stratum for certainty selections, so that there were 25 explicit strata in total. Variables used for implicit stratification and the respective number of levels can also be found in Table 4.1.

As the sampling frame was always finally sorted by school size, school size was also an implicit stratification variable, though it is not listed in Table 4.1. The use of school size as an implicit stratification variable provides a degree of control over the student sample size so as to possibly avoid the sampling of too many relatively large schools or too many relatively small schools.

**Table 4.1 Stratification variables used in PISA 2015 [Part 1/3]**

Country/economy	Explicit stratification variables	Number of explicit strata	Implicit stratification variables
Albania	Urbanisation (2); Geographical division (3); Funding (2); Certainty selections	13	ISCED level (3)
Algeria	Region (4); Urbanisation (3)	12	ISCED level (4); School gender composition (3)
Argentina	Region (6)	6	Funding (2); Education level (4); Urbanisation (2); Secular/Religious (2)
Australia	State/Territory (8); Sector (3); Modal grade (2); Certainty selections	49	Urbanisation (3); School gender composition (3); School socioeconomic level (11); ISCED level (3)
Austria	AUT/Oberoesterreich (2); Programme – for rest of Austria only (17); Oberoesterreich programme group (8); Certainty selections	26	School Type (3); Region (9); OOE programme (18); Percentage of females within programmes (118)
Belgium	Region (3); Form of education – Flanders (5), French Community (3), German Community (2); Funding – for Flanders only (2); ISCED level (3), Educational tracks – for French Community only (4)	32	Type of school--for French Community only (4); Grade repetition (5), Percentage of females (4)
Brazil	State (27); Modal grade (2); Certainty selections	55	Funding (5); HDI quintiles (5); ISCED level (3); Capital/Interior (2); Urbanisation (2)
Bulgaria	Region (11)	11	Type of school (8); Size of settlement (5)
Canada	Province (10); Language (3); School size (7); Certainty selections	98	Urbanisation (3); Funding (2); ISCED level (3)
Chile	Funding (3); School level (3); School track (4); Certainty selections	25	National test score level (3); Percentage of females (6); Urbanisation (2); Region (4)
B-S-J-G (China)*	Area of Beijing--for Beijing only (2); Urbanisation (3); ISCED programme orientation (2); ISCED level (3)	53	Selectivity (3); Funding (2)
Colombia	Region (6); Modal grade (2); Main shift (2); Certainty selections	23	Urbanisation (2); Funding (2); Weekend school or not (2); School gender composition (5); ISCED programme orientation (4)
Costa Rica	School type (5); Certainty selections	6	School track (2); Urbanisation (2); Shift (2); Region (27); ISCED level (3)
Croatia	Dominant programme type (6); Certainty selections	7	School gender composition (3); Urbanisation (3); Region (6)
Cyprus <sup>1</sup>	ISCED programme orientation (3); Funding (2); Urbanisation (2)	8	Language (2); ISCED level (3)
Czech Republic	Programmes (6); Region for programmes 1 and 2 (14)	32	School size (3); Region for programmes 3, 4, 5 (14); School gender composition (3)
Denmark	Immigrant levels (5); Certainty selections	6	School type (7); ISCED level (3); Urbanisation (5); Region (5); FO group (3)
Dominican Republic	Funding (3); Urbanisation (2); ISCED level (3); Modal grade (2); Certainty selections	18	Shift (6); School size (4); Programme (3)
Estonia	Language (3); Certainty selections	4	School type (3); Urbanisation (2); County (15); Funding (2)



Table 4.1 Stratification variables used in PISA 2015 [Part 2/3]

Country/economy	Explicit stratification variables	Number of explicit strata	Implicit stratification variables
Finland	Region (5); Urbanisation (2)	10	Regional state administrative agencies – for major regions of Northern & Eastern Finland and Swedish-speaking regions only (6); School type (7)
France	School type (4) only for non-small schools; School size (3)	6	Funding (2)
Georgia	Region (12); Funding (2)	23	Language (11)
FYROM	ISCED level (2); Orientation (3)	4	Urbanisation (2)
Germany	School category (3); State – for normal schools only (16)	18	State – for other schools only (16); School type – for normal schools only (5)
Greece	Urbanisation (3)	3	Funding and region (16); School type (3)
Hong Kong (China)	Funding (4); Modal grade (2)	5	Student Academic Intake (4)
Hungary	School type (6)	6	Region (7); Mathematics performance (6)
Iceland	Region (9); School size (4)	32	Urbanisation (2)
Indonesia	National examination result (3)	3	Funding (2); School type (3); Region (8)
Ireland	School Size (3); School type (3)	9	Socioeconomic quartile (4); School gender composition (4)
Israel	School type (12)	12	ISCED level (3); School size (2); Socioeconomic status (3); District (2)
Italy	Region (13); Study programme (5); Certainty selections	65	Region (10) for "Rest of Italy" stratum; Funding (2)
Japan	Funding (2); Orientation (2)	4	Levels of proportion of students taking university/college entrance exams (4)
Jordan	School type / Funding (6)	6	Urbanisation (2); School gender composition (3); Level (2); Shift (2)
Kazakhstan	Region – for non-intellectual schools only (15); Language – for non-intellectual schools only (3); Intellectual school or not (2)	49	Region – for intellectual schools only (13); Urbanisation (2); ISCED level (3); ISCED programme orientation (2); Funding (2)
Korea	School level (2); Orientation (2)	3	Urbanisation (3); School gender composition (3)
Kosovo	Region (7); Urbanisation (2); Certainty selections	15	Study programme (4)
Latvia	Urbanisation (4); Certainty selections	5	School type/level (5)
Lebanon	ISCED level (5); Funding (2); Urbanisation (2); Certainty selections	13	School language (3); School gender composition (3)
Lithuania	School language (3); Urbanisation – for Lithuanian language schools only (4); School type – for Lithuanian language schools (5); Certainty selections	25	School language for "multi-language stratum" (4); Urbanisation – for non-Lithuanian language schools (4); School type – for non-Lithuanian language schools (5); Funding (2)
Luxembourg	School type (6)	6	School gender composition (3)
Macao (China)	School type (3); Study programme (2); Language (5)	10	School gender composition (3); Secular or religious (2)
Malaysia	School category (6); State – except for MOE Fully-Residential Schools (4)	9	School type (16); Urbanisation (2); School gender composition (3); ISCED level (2)
Malta	School management (3); Study programme – for state schools only (7)	9	School gender composition (3)
Mexico	School level (2); School size (3)	6	School programme (7); Funding (2); Urbanisation (2)
Moldova	Language (3); Urbanisation (3); ISCED level (3)	27	Funding (2); Study programme (6)
Montenegro	Programme (4); Region (3)	11	School gender composition (3)
Netherlands	School track (3)	3	Programme category (10)
New Zealand	School size (3); Certainty selections	4	School decile (4); Funding (2); School gender composition (3); Urbanisation (2)
Norway	School level (3)	3	None
Peru	Funding (2); Urbanisation (2); Modal grade (2)	8	Region (26); School gender composition (3); School type (6)
Poland	School type (3)	3	Vocational school or not (2); Funding (2); Locality (4); School gender composition (3)
Portugal	Geographic region (25); Modal grade (2)	50	Funding (2); Urbanisation (3); ISCED programme orientation (3)
Puerto Rico (USA) <sup>2</sup>	Funding (2)	2	Grade span (5); District (8); Urbanisation (5)
Qatar	School type (6)	6	School gender composition (3); Language (2); Level (5); Funding (2); ISCED programme orientation (3)
Romania	Programme (2)	2	Language (3); Urbanisation (2); LIC type (3)
Russian Federation	Region (42)	42	Location/Urbanisation (9); School type (3)
Scotland	Funding (2); School attainment (6)	7	School gender composition (3); Area type (6)
Singapore	Funding (2); School level (2); Certainty selections	4	School gender composition (3)

**Table 4.1 Stratification variables used in PISA 2015 [Part 3/3]**

Country/economy	Explicit stratification variables	Number of explicit strata	Implicit stratification variables
Slovak Republic	School type (3); Region (3)	9	Sub-region (8); School type (7); Language (3); Exam (10); ESCS (7); Funding (3); Grade repetition level (163)
Slovenia	Programme/Level (7)	7	Location/Urbanisation (5); School gender composition (3)
Spain	Region (18); Funding (2); Linguistic model – for the Basque region only (3); Certainty selections	41	none
Sweden	Funding (2); ISCED level (2); Urbanisation (3)	8	Geographic LAN – for upper secondary only (21); Responsible authority – for upper secondary only (3); Level of immigrants – for lower secondary/mixed only (3); Income Quartiles – for lower secondary/mixed only (4)
Switzerland	Language (3); ISCED level (3); Funding (3); Certainty selections	25	School type (22); Canton (26)
Chinese Taipei	School type (6); Funding (2); Certainty selections	13	Region (6); School gender composition (3)
Thailand	Administration (7); ISCED level (3)	16	Region (9); Urbanisation (2); School gender composition (3)
Trinidad and Tobago	Educational districts (8); Management (3)	22	School gender composition (3); Urbanisation (2)
Tunisia	Geographical area (6); Urbanisation (3)	18	ISCED level (3); Funding (2); Percentage of repeaters (4)
Turkey	Region (12); Programme type (4)	36	School type (10); School gender composition (3); Urbanisation (2); Funding (2)
United Arab Emirates	Emirate (7); Curriculum (5); Funding (2); Certainty selections	43	School gender composition (3); Language (2); ISCED level (3); ISCED programme orientation (2)
United Kingdom	Country (3); School type (9); Region (12), Modal grade – England only (2); School gender composition (3); Certainty selections	96	School performance – England and Wales only (6); Local authority (204)
United States	Region (4); Funding (2); Public school, no modal grade (1)	9	Grade span (5); Urbanisation (4); Minority Status (2); School gender composition (3); State (51)
Uruguay	Institutional sector (4); School level (3); Certainty selections	11	Location/Urbanisation (4); School gender composition (3)
Viet Nam	Geographical zone (3); Funding (2); Urbanisation (3)	15	Region (6); Province (63); School type (5); Study commitment (2)

\* B-S-J-G (China) refers to the four PISA-participating China provinces: Beijing, Shanghai, Jiangsu and Guangdong.

1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

2. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

## Assigning a measure of size to each school

For the probability proportional to size sampling method used for PISA, a Measure of Size (MOS) derived from *ENR* was established for each school on the sampling frame. *MOS* was generally constructed as:  $MOS = \max(ENR, TCS)$ . This differed slightly in the case of small schools treatment, discussed later.

Thus, the measure of size was equal to the enrolment estimate (*ENR*), unless enrolment was less than the *TCS*, in which case the measure of size was set equal to the target cluster size. In most countries, the *MOS* was equal to *ENR* or the *TCS*, whichever was larger.

As schools were sampled with probability proportional to size, setting the measure of size of small schools to 42 students (or 35 for paper-based countries) was equivalent to drawing a simple random sample of small schools. That is, small schools would have an equally likely chance of being selected to participate. However, please see the "Treatment of small schools" for details on how small schools were sampled.

## School sample selection

### School sample allocation over explicit strata

The total number of schools to be sampled in each country needed to be allocated among the explicit strata so that the expected proportion of students in the sample from each explicit stratum was approximately the same as the population proportions of PISA-eligible students in each corresponding explicit stratum. There were two exceptions. If very small schools required under-sampling, students in them had smaller percentages in the sample than in the population. To compensate for the resulting loss of sample, the large schools had slightly higher percentages in the sample than the corresponding population percentages. The other exception occurred if only one school was allocated to any explicit stratum. In this case, two schools were allocated for selection in the stratum to aid with variance estimation.



### **Sorting the sampling frame**

The *School Sampling Preparation Manual* indicated that, prior to selecting the school sample, schools in each explicit stratum were to be sorted by a limited number of variables chosen for implicit stratification and finally by the *ENR* value within each implicit stratum. The schools were first to be sorted by the first implicit stratification variable, then by the second implicit stratification variable within the levels of the first implicit stratification variable, and so on, until all implicit stratification variables were used. This gave a cross-classification structure of cells, where each cell represented one implicit stratum on the school sampling frame. The sort order was alternated between implicit strata, from high to low and then low to high, etc., through all implicit strata within an explicit stratum.

### **Determining which schools to sample**

The PPS-systematic sampling method used in PISA first required the computation of a sampling interval for each explicit stratum. This calculation involved the following steps:

- recording the total measure of size,  $S$ , for all schools in the sampling frame for each specified explicit stratum
- recording the number of schools,  $D$ , to be sampled from the specified explicit stratum, which was the number allocated to the explicit stratum
- calculating the sampling interval,  $I$ , as follows:  $I = S/D$
- including in the sample all schools for which the school's size measure exceed  $I$  (known as certainty schools)
- removing certainty schools from the frame, recalculating  $S$ ,  $D$ , and  $I$
- recording the sampling interval,  $I$ , to four decimal places.

Next, a random number had to be generated for each explicit stratum. The generated random number (*RN*) was from a uniform distribution between zero and one and was to be recorded to four decimal places.

The next step in the PPS selection method in each explicit stratum was to calculate selection numbers – one for each of the  $D$  schools to be selected in the explicit stratum. Selection numbers were obtained using the following method:

- Obtaining the first selection number by multiplying the sampling interval,  $I$ , by the random number, *RN*. This *RN* number is a random number between zero and one, and to 4 decimal places. This first selection number was used to identify the first sampled school in the specified explicit stratum.
- Obtaining the second selection number by adding the sampling interval,  $I$ , to the first selection number. The second selection number was used to identify the second sampled school.
- Continuing to add the sampling interval,  $I$ , to the previous selection number to obtain the next selection number. This was done until all specified line numbers (1 through  $D$ ) had been assigned a selection number.

Thus, the first selection number in an explicit stratum was  $RN \times I$ , the second selection number was  $(RN \times I) + I$ , the third selection number was  $(RN \times I) + I + I$ , and so on.

Selection numbers were generated independently for each explicit stratum, with a new random number generated for each explicit stratum.

### **Identifying the sampled schools**

The next task was to compile a cumulative measure of size in each explicit stratum of the school sampling frame that assisted in determining which schools were to be sampled. Sampled schools were identified as follows:

Let  $Z$  denote the first selection number for a particular explicit stratum. It was necessary to find the first school in the sampling frame where the cumulative *MOS* equalled or exceeded  $Z$ . This was the first sampled school. In other words, if  $C_s$  was the cumulative *MOS* of a particular school  $S$  in the sampling frame and  $C_{(s-1)}$  was the cumulative *MOS* of the school immediately preceding it, then the school in question was selected if  $C_s$  was greater than or equal to  $Z$ , and  $C_{(s-1)}$  was strictly less than  $Z$ . Applying this rule to all selection numbers for a given explicit stratum generated the original sample of schools for that stratum.

#### Box 4.1 Illustration of probability proportional to size (PPS) sampling

To illustrate these steps, suppose that in an explicit stratum in a participant country, the PISA-eligible student population is 105 000, then:

- the total measure of size,  $S$ , for all schools is 105 000
- the number of schools,  $D$ , to be sampled is 150
- calculating the sampling interval,  $I$ ,  $105\,000/150 = 700$
- generate a random number,  $RN$ , 0.3230
- the first selection number is  $700 \times 0.3230 = 226$  and it was used to identify the first sampled school in the specified explicit stratum
- the second selection number is  $226 + 700 = 926$  and it was used to identify the second sampled school
- the third selection number is  $926 + 700 = 1\,626$  and it was used to identify the third sampled school, and so on until the end of the school list is reached.

This will result in a school sample size of 150 schools.

The table below also provides these example data. The school that contains the generated selection number within its cumulative enrolment is selected for participation.

School	MOS	Cumulative MOS ( $C_s$ )	Selection number	School selection
001	550	550	226	Selected
002	364	914		
003	60	974	926	Selected
004	93	1 067		
005	88	1 155		
006	200	1 355		
007	750	2 105	1 626	Selected
008	72	2 177		
009	107	2 284		
010	342	2 626	2 326	Selected
011	144	2 770		
...	...	...	...	...

#### **Identifying replacement schools**

Each sampled school in the main survey was assigned two replacement schools from the school sampling frame, if possible, identified as follows: for each sampled school, the schools immediately preceding and following it in the explicit stratum, which was ordered within by the implicit stratification, were designated as its replacement schools. The school immediately following the sampled school was designated as the first replacement and labelled  $R_1$ , while the school immediately preceding the sampled school was designated as the second replacement and labelled  $R_2$ . The *School Sampling Preparation Manual* noted that in small countries, there could be problems when trying to identify two replacement schools for each sampled school. In such cases, a replacement school was allowed to be the potential replacement for two sampled schools (a first replacement for the preceding school, and a second replacement for the following school), but an actual replacement for only one school. Additionally, it may have been difficult to assign replacement schools for some very large sampled schools because the sampled schools appeared close to each other in the sampling frame. There were times when it was only possible to assign a single replacement school, or even none, when two consecutive schools in the sampling frame were sampled. That is, no unsampled schools existed between sampled schools.

Exceptions were allowed if a sampled school happened to be the last school listed in an explicit stratum. In this case the two schools immediately preceding it were designated as replacement schools. Similarly, for the first school listed in an explicit stratum, the two schools immediately following it were designated as replacement schools.

#### **Assigning school identifiers**

To keep track of sampled and replacement schools in the PISA database, each was assigned a unique, three-digit school code sequentially numbered starting with one within each explicit stratum (each explicit strata was numbered with



a separate two-digit stratum code). For example, if 150 schools are sampled from a single explicit stratum, they are assigned identifiers from 001 to 150. First replacement schools in the main survey are assigned the school identifier of their corresponding sampled schools, incremented by 300. For example, the first replacement school for sampled school 023 is assigned school identifier 323. Second replacement schools in the main survey are assigned the school identifier of their corresponding sampled schools, but incremented by 600. For example, the second replacement school for sampled school 136 took the school identifier 736.

### **Tracking sampled schools**

NPMs were encouraged to make every effort to confirm the participation of as many sampled schools as possible to minimise the potential for non-response biases. Each sampled school that did not participate was replaced if possible. NPMs contacted replacement schools only after all contacts with sampled schools were made. If the unusual circumstance arose whereby both an original school and a replacement participated, only the data from the original school were included in the weighted data, provided that at least 50% of the PISA-eligible, non-excluded students had participated. If this was not the case, it was permissible for the original school to be labelled as a nonrespondent and the replacement school as the respondent, provided that the replacement school had at least 50% of the PISA-eligible, non-excluded students as participants.

## **Special school sampling situations**

### **Treatment of small schools**

In PISA, schools were classified as very small, moderately small or large. A school was classified as large if it had an *ENR* above the *TCS* (42 students in most countries). A moderately small school had an *ENR* in the range of one-half the *TCS* to *TCS* (21 to 41 students in most countries). A very small school had an *ENR* less than one-half the *TCS* (20 students or fewer in most countries). Schools with especially few students were further classified as either very small schools with an *ENR* of zero, one, or two students or very small schools with an *ENR* greater than two students but less than one-half the *TCS*. Unless they received special treatment in the sampling, the occurrence of small schools in the sample will reduce the sample size of students for the national sample to below the desired target because the within-school sample size would fall short of expectations. A sample with many small schools could also be an administrative burden with many testing sessions with few students. To minimise these problems, procedures were devised for managing small schools in the sampling frame.

To balance the two objectives of selecting an adequate sample of small schools but not too many small schools so as to hurt student yield, a procedure was recommended that assumed the underlying idea of under-sampling the very small schools by a factor of two (those with an *ENR* greater than two but less than one-half the *TCS*) and under-sampling the very small schools with zero, one, or two students by a factor of four and to proportionally increasing the number of large schools to sample. To determine whether very small schools should be undersampled and if the sample size needed to be increased to compensate for small schools, the following test was applied.

- If the percentage of students in very small schools ( $ENR < TCS/2$ ) was 1 percent or MORE, then very small schools were undersampled and the school sample size increased, sufficient to maintain the required overall yield.
- If the percentage of students in very small schools ( $ENR < TCS/2$ ) was LESS than 1 percent, and the percentage of students in moderately small schools ( $TCS/2 < ENR < TCS$ ) was 4 percent or MORE, then there was no required undersampling of very small schools but the school sample size was increased, sufficient to maintain the required overall yield.

If none of these conditions were true, then the small schools contained such a small proportion of the PISA population that they were unlikely to reduce the sample below the desired target. In this case, no undersampling of very small schools was needed nor an increase to the school sample size to compensate for small schools.

Building on the PISA 2012 treatment of small schools, the PISA 2015 approach added to the criteria for undersampling very small schools by including the condition where the percentage of schools on the frame that are the very smallest (*ENR* of zero, one, or two) is 20 percent or more. This modification was for the infrequent situation where very small schools ( $ENR < TCS/2$ ) overall contain less than 1 percent of total frame enrolment while at the same time these very smallest schools account for a large percentage of total schools on the frame. If this condition was met and no undersampling was otherwise required based on the percentage of enrolment in very small schools, very small schools were undersampled to avoid having too many of these in the school sample. Even though undersampling can reduce the number of these

in the sample from what could be expected without undersampling, when very small schools account for such a large percentage of schools on the frame it is likely that a relatively large number of them (but not a large proportion) will be selected. A minor increase to the sample size was needed in this case to safeguard the needed student sample size.

If the number of very small schools was to be controlled in the sample without creating explicit strata for these small schools, this was accomplished by assigning a measure of size (*MOS*) of  $TCS/2$  to those very small schools with an *ENR* greater than two but less than  $TCS/2$  and a measure of size equal to the  $TCS/4$  for the very small schools with an *ENR* of zero, one, or two. In effect, very small schools with a measure of size equal to  $TCS/2$  were under-sampled by a factor of two (school probability of selection reduced by half), and the very small schools with a measure of size equal to  $TCS/4$  were under-sampled by a factor of four (school probability of selection reduced by three-fourths). This was accomplished as follows and was a standard procedure followed in all countries.

The formulae below assume an initial target school sample size of 150 and a target student sample size of 6 300.

- Step 1: From the complete sampling frame, find the proportions of total *ENR* that come from very small schools with *ENR* of zero, one or two ( $P1$ ), very small schools with *ENR* greater than two but fewer than  $TCS/2$  ( $P2$ ), moderately small schools ( $Q$ ) and large schools ( $R$ ). Thus,  $P1 + P2 + Q + R = 1$ .
- Step 2: Calculate the value  $L$ , where  $L = 1.0 + 3(P1)/4 + (P2)/2$ . Thus  $L$  is a positive number slightly more than 1.0.
- Step 3: The minimum sample size for large schools is equal to  $150 \times R \times L$ , rounded up to the nearest integer. It may need to be enlarged because of national considerations, such as the need to achieve minimum sample sizes for geographic regions or certain school types.
- Step 4: Calculate the mean value of *ENR* for moderately small schools (*MENR*), and for very small schools (*V1ENR* and *V2ENR*). *MENR* is a number in the range of  $TCS/2$  to  $TCS$ , *V2ENR* is a number larger than two but no greater than  $TCS/2$ , and *V1ENR* is a number in the range of zero to two.
- Step 5: The number of schools that must be sampled from the moderately small schools is given by:  $(6\ 300 \times Q \times L)/(MENR)$ .
- Step 6: The number of schools that must be sampled from the very small schools (type  $P2$ ) is given by:  $(3\ 150 \times P2 \times L)/(V2ENR)$ .
- Step 7: The number of schools that must be sampled from the very small schools (type  $P1$ ) is given by:  $(1\ 575 \times P1 \times L)/(V1ENR)$ .

To illustrate the steps, suppose that in a participant country, the *TCS* is equal to 42 students, with 10% of the total enrolment of 15 year olds in moderately small schools, and 5% in each type of very small schools,  $P1$  and  $P2$ . Suppose that the average enrolment in moderately small schools is 25 students, in very small schools (type  $P2$ ) it is 12 students, and in very small schools (type  $P1$ ) it is 1.5 students.

- Step 1: The proportions of total *ENR* from very small schools is  $P1 = 0.05$  and  $P2 = 0.05$ , from moderately small schools is  $Q = 0.1$ , and from large schools is  $R = 0.8$ . The proportion of the very smallest schools on the frame was not more than 20%. It can be shown that  $0.05 + 0.05 + 0.1 + 0.8 = 1.0$ .
- Step 2: Calculate the value  $L$ .  $L = 1.0 + 3(0.05)/4 + (0.05/2)$ . Thus  $L = 1.0625$ .
- Step 3: The minimum sample size for large schools is equal to  $150 \times 0.8 \times 1.0625 = 127.5$ . That is, at least 128 (rounded up to the nearest integer) of the large schools must be sampled.
- Step 4: The mean value of *ENR* for moderately small schools (*MENR*) is given in this example as 25, very small schools of type  $P2$  (*V2ENR*) as 12, and very small schools of type  $P1$  (*V1ENR*) as 1.5.
- Step 5: The number of schools that must be sampled from the moderately small schools is given by  $(6\ 300 \times 0.1 \times 1.0625)/25 = 26.8$ . At least 27 (rounded up to the nearest integer) moderately small schools must be sampled.
- Step 6: The number of schools that must be sampled from the very small schools (type  $P2$ ) is given by  $(3\ 150 \times 0.05 \times 1.0625)/12 = 13.9$ . At least 14 (rounded up to the nearest integer) very small schools of type  $P2$  must be sampled.
- Step 7: The number of schools that must be sampled from the very small schools (type  $P1$ ) is given by  $(1\ 575 \times 0.05 \times 1.0625)/1.5 = 55.8$ . At least 56 (rounded up to the nearest integer) very small schools of type  $P1$  must be sampled.



Combining these different sized school samples gives a total sample size of  $128 + 27 + 14 + 56 = 225$  schools. Before considering school and student non-response, the larger schools will yield an initial sample of approximately  $128 \times 42 = 5\,376$  students. The moderately small schools will give an initial sample of approximately  $27 \times 25 = 675$  students, very small schools of type  $P_2$  will give an initial sample size of approximately  $14 \times 12 = 168$  students, and very small schools of type  $P_1$  will give an initial sample size of approximately  $56 \times 1.5 = 84$  students. The total expected sample size of students is therefore  $5\,376 + 675 + 168 + 84 = 6\,303$ .

This procedure, called small school analysis, was done not just for the entire school sampling frame, but for each individual explicit stratum. An initial allocation of schools to explicit strata provided the starting number of schools and students to project for sampling in each explicit stratum. The small school analysis for a single unique explicit stratum indicated how many very small schools of each type (assuming under-sampling, if needed), moderately small schools and large schools would be sampled in that stratum. Together, these provided the final sample size,  $n$ , of schools to select in the stratum. Based on the stratum sampling interval and random start, large, moderately small, and very small schools were sampled in the stratum, to a total of  $n$  sampled schools. Because of the random start, it was possible to have more or less than expected of the very small schools of either type,  $P_1$  or  $P_2$ , of the moderately small schools, and of the large schools. The total number of sampled schools however was fixed at  $n$ , and the number of expected students to be sampled was always approximate to what had been projected from the unique stratum small school analysis.

### **PISA and national study overlap control**

The main studies for PISA 2015 and a national (non-PISA) survey were to occur at approximately the same time in some participating countries. Because of the potential for increased burden, an overlap control procedure was used for seven countries (Canada (TIMSS), Hong Kong (China) (TIMSS), Ireland (TIMSS), Norway (TIMSS), Sweden (TIMSS), United Kingdom (TIMSS), and Mexico's national option state sample (Mexico's 2015 national sample)) who requested that there be a minimum incidence of the same schools being sampled for both PISA and their national (non-PISA) study. This overlap control procedure required that the same school identifiers be used on the PISA and the national study school frames for the schools in common across the two assessments.

The national study samples were usually selected before the PISA samples. Thus, for countries requesting overlap control, the national study centre supplied the international contractor with their school frames, national school IDs, each school's probability of selection, and an indicator showing which schools had been sampled for the national study.

Sample selections for PISA and the national study could totally avoid overlap of schools if schools which would have been selected with high probability for either study had their selection probabilities capped at 0.5. Such an action would make each study's sample slightly less than optimal, but this might be deemed acceptable when weighed against the possibility of low response rates due to the burden of participating in two assessments. Only Hong Kong (China) requested this for PISA 2015. Therefore, if any schools had probabilities of selection greater than 0.5 on either study frame for the other countries where overlap control was implemented, these schools had the possibility to be selected to be in both studies.

To control overlap of schools between PISA and another sample, the sample selection of schools for PISA adopted a modification of an approach due to Keyfitz (1951) based on Bayes Theorem. To use PISA and TIMSS (an international study controlled for with the Keyfitz method during the 2009 PISA) in an example of the overlap control approach to minimise overlap, suppose that  $PROBP$  is the PISA probability of selection and  $PROBI$  is the ICCS probability of selection. Then a conditional probability of a school's selection into PISA ( $CPROB$ ) is determined as follows:

**4.1**

$$CPROB = \begin{cases} \max \left[ 0, \left( \frac{PROBI + PROBP - 1}{PROBI} \right) \right] & \text{if the school was a TIMSS school} \\ \min \left[ 1, \frac{PROBP}{(1 - PROBI)} \right] & \text{if the school was not a TIMSS school} \\ PROBP & \text{if the school was not a TIMSS eligible school} \end{cases}$$

Then a conditional CMOS variable was created to coincide with these conditional probabilities as follows:

$$CMOS = CPROB \times \text{stratum sampling interval}$$

The PISA school sample was then selected using the line numbers created as usual (see earlier section), but applied to the cumulated CMOS values (as opposed to the cumulated MOS values). Note that it was possible that the resulting PISA sample size could be slightly lower or higher than the originally assigned PISA sample size, but this was deemed acceptable.

## Monitoring school sampling

PISA 2015 Technical Standard 1.13 states that, as in the previous cycles, the international contractor should select the school samples unless otherwise agreed upon (see Appendix F). Japan was the only participant that selected their own school sample, doing so for reasons of confidentiality.

Sample selection for Japan was replicated by the international contractor using the same random numbers as used by the Japanese national centre, to ensure quality in this case. All other participating countries' school samples were selected by and checked in detail by the international contractor. To enable this, all countries were required to submit sampling information on forms associated with the following various sampling tasks:

- time of testing and age definition for both the field trial and main study were captured on Sampling Task 1 (see below) at the time of the field trial, with updates being possible before the main study
- information about stratification for the field trial and for the main study was recorded on Sampling Task 2
- forms or data associated with Sampling Tasks 3, 4, 5 and 6 were all for the field trial
- the national desired target population information for the main study was captured on the form associated with Sampling Task 7a
- information about the defined national target population was recorded on the form associated with Sampling Task 7b;
- the description of the sampling frame was noted on the form associated with Sampling Task 8a
- the school sampling frame was created in one spreadsheet and the list of any excluded schools in a second spreadsheet associated with Sampling Task 8b.

The international contractor completed school sampling and, along with the school sample, returned other information (small school analyses, school allocation, and a spreadsheet that countries could use for tracking school participation). Table 4.2 provides a summary of the information required for each sampling task and the timetables (which depended on national assessment periods).

**Table 4.2 Schedule of school sampling activities**

Activity	Submit to Consortium	Due Date
Update time of testing and age definition of population to be tested	Sampling Task 1 – time of testing and age definition	Update what was submitted at the time of the FT, two months before the school sample is to be selected
Finalise explicit and implicit stratification variables	Sampling Task 2 – stratification and other information	Update what was submitted at the time of the FT, two months before the school sample is to be selected
Define national desired target population	Sampling Task 7a – national desired target population	Submit two months before the school sample is to be selected
Define national defined target population	Sampling Task 7b – national defined target population	Submit two months before the school sample is to be selected
Create and describe sampling frame	Sampling Task 8a – sampling frame description	Submit two months before the school sample is to be selected
Submit sampling frame	Sampling Task 8b – sampling frame (in one Excel® sheet), and excluded schools (in another Excel® sheet)	Submit two months before the school sample is to be selected
Decide how to treat small schools	Treatment of small schools	The international contractor will complete and return this information to the NPM about one month before the school sample is to be selected
Finalise sample size requirements	Sampling Task 9 – sample allocation by explicit strata	The international contractor will complete and return this information to the NPM about one month before the school sample is to be selected
Describe population within strata	Population counts by strata	The international contractor will complete and return this information to the NPM when the school sample is sent to the NPM
Select the school sample	Sampling Task 10 – school sample selection	The international contractor will return the sampling frame to the NPM with sampled schools and their replacement schools identified and with PISA IDs assigned when the school sample is selected
Review and agree to the sampling form required as input to KeyQuest	Sampling Task 11 – reviewing and agreeing to the Sampling Form for KeyQuest (SFKQ)	Countries had one month after their sample was selected to agree to their SFKQ
Submit sampling data	Sampling Task 12 – school participation information and data validity checks	Submit within one month of the end of the data collection period



Once received from each participating country, each set of information was reviewed and feedback was provided to the country. Forms were only approved after all criteria were met. Approval of deviations was only given after discussion and agreement by the international contractors. In cases where approval could not be granted, countries were asked to make revisions to their sample design and sampling forms and resubmit.

Checks that were performed when monitoring each sampling task follow. Although all sampling tasks were checked in their entirety, the below paragraphs contain matters that were explicitly examined.

Just after countries submitted their main survey sampling tasks, the international contractor verified all special situations known in each participating country. Such special situations included whether or not: the TCS value differed from 42 or 35 students; the Financial Literacy Assessment was being conducted; the Teacher Questionnaire was being conducted; overlap control procedures with a national (non-PISA) survey were required; there was any regional or other type of oversampling; the UH booklet would be used; and any grade or other type of student sampling would be used. Additionally, any countries with fewer than 4 500 or just over 4 500 assessed students in either PISA 2009 or 2012 had increased school sample sizes discussed and agreed upon. Additionally, countries which had too many PISA 2012 exclusions were warned about not being able to exclude any schools in the field for PISA 2015. Finally, any countries with effective student sample sizes less than 400 in PISA 2012 also had increased school sample sizes discussed and agreed upon.

### ***Sampling task 0: Languages of instruction***

The ST0 was a new task for PISA 2015. The information collected was not new but used to be collected as part of the ST2. Language information was needed much earlier in the cycle for PISA 2015 so this new task was created for its collection.

- Language distributions were compared with those of PISA 2012 for countries which had participated in PISA 2012. Differences in languages and/or the percentage distribution were queried.
- The existence of international/foreign schools was asked about.
- Checks were done on the appropriate inclusion of languages in the FT along with proper verification plans.
- Languages which were planned for MS exclusion were scrutinised.

### ***Sampling task 1: Time of testing and age definition***

- Assessment dates had to be appropriate for the selected target population dates.
- Assessment dates could not cover more than a 42-day period unless agreed upon.
- Assessment dates could not be within the first six weeks of the academic year.
- If assessment end dates were close to the end of the target population birth date period, NPMs were alerted not to conduct any make-up sessions beyond the date when the population birth dates were valid.

### ***Sampling task 2: Stratification (and other information)***

- Each participating country used explicit strata to group similar schools together to reduce sampling variance and to ensure representativeness of students in various school types using variables that might be related to outcomes. The international contractor assessed each country's choice of explicit stratification variables. If a country was known to have school tracking or distinct school programmes and these were not among the explicit stratification variables, a suggestion was made to include this type of variable.
- Dropping variables or reducing levels of stratification variables used in the past was discouraged and only accepted if the National Centre could provide strong reasons for doing so.
- Adding variables for explicit stratification was encouraged if the new variables were particularly related to outcomes. Care was taken not to have too many explicit strata though.
- Levels of variables and their codes were checked for completeness.
- If no implicit stratification variables were noted, suggestions were made about ones that might be used. In particular, if a country had single gender schools and school gender was not among the implicit stratification variables, a suggestion was made to include this type of variable to ensure no sample gender imbalances. Similarly, if there were ISCED school level splits, the ISCED school level was also suggested as an explicit or implicit stratification variable.
- Without overlap control there is nearly as good control over sample characteristics compared to population characteristics whether explicit or implicit strata are used. With overlap control some control is lost when using

implicit strata, but not when using explicit strata. For countries which wanted overlap control with a national non-PISA survey, as many as possible of their implicit stratification variables were made explicit stratification variables.

- If grade or other national option sampling, or special oversampling of subpopulations of PISA students were chosen options, checks were done to ensure there was only one student sampling option per explicit stratum.

#### **Sampling task 7a: National desired target population**

- The total national number of 15 year olds of participating countries was compared with those from previous cycles. Differences, and any kind of trend, were queried.
- Large deviations between the total national number of 15 year olds and the enrolled number of 15 year olds were questioned.
- Large increases or decreases in enrolled population numbers compared to those from previous PISA cycles were queried, as were increasing or decreasing trends in population numbers since PISA 2000.
- Any population to be omitted from the international desired population was noted and discussed, especially if the percentage of 15 year olds to be excluded was more than 0.5% or if it was substantially different or not noted for previous PISA cycles.
- Calculations did not have to be verified as in previous cycles as such data checks were built into the form.
- For any countries using a three-stage design, a Sampling Task 7a form also needed to be completed for the full national desired population as well as for the population in the sampled regions.
- For countries having adjudicated regions, a Sampling Task 7a form was needed for each region.
- Data sources and the year of the data were required. If websites were provided with an English page option, the submitted data was verified against those sources.

#### **Sampling task 7b: National defined target population**

- The population value in the first question needed to correspond with the final population value on the form for Sampling Task 7a. This was accomplished through built-in data checks.
- Reasons for excluding schools for reasons other than special education needs were checked for appropriateness (i.e. some operational difficulty in assessing the school). In particular, school-level language exclusions were closely examined to check correspondence with what had been noted about language exclusions on Sampling Task 0.
- Exclusion types and extents were compared to those recorded for PISA 2012 and previous cycles. Differences were queried.
- The number and percentage of students to be excluded at the school level and whether the percentage was less than the guideline for maximum percentage allowed for such exclusions were checked.
- Reasonableness of assumptions about within-school exclusions was assessed by checking previous PISA coverage tables. If there was an estimate noted for “other”, the country was queried for reasonableness about what the “other” category represented. If it was known the country had schools where some of the students received instruction in minority languages not being tested, an estimate for the within-school exclusion category for “no materials available in the student’s language of instruction” was necessary.
- Form calculations were verified through built-in data checks, and the overall coverage figures were assessed.
- If it was noted that there was a desire to exclude schools with only one or two PISA-eligible students at the time of contact, then the school sampling frame was checked for the percentage of population that would be excluded. If countries had not met the 2.5% school-exclusion guideline and if these schools would account for not more than 0.5% and if within-school exclusions looked similar to the past and were within 2.5%, then the exclusion of these schools at the time of contact was agreed upon with the understanding that such exclusion not cause entire strata to be missing from the student data.
- The population figures on this form after school-level exclusions were compared against the aggregated school sampling frame enrolment. School-level exclusion totals also were compared to those tabulated from the excluded school sheet of the Sampling frame, ST8b. Differences were queried.
- For any countries using a three-stage design, a Sampling Task 7b form also needed to be completed for the full national defined population as well as for the population in the sampled regions.



- For countries having adjudicated regions, a Sampling Task 7b form was needed for each region.
- Data sources and the year of the data were required. If websites were provided with an English page option, the submitted data was verified against those sources.

#### ***Sampling task 8a: Sampling frame description***

- Special attention was given to countries who reported on this form that a three-stage sampling design was to be implemented and additional information was sought from countries in such cases to ensure that the first-stage sampling was done adequately.
- The type of school-level enrolment estimate and the year of data availability were assessed for reasonableness.
- Countries were asked to provide information for each of various school types,<sup>2</sup> whether those schools were included on or excluded from the sampling frame, or the country did not have any of such schools. The information was matched to the different types of schools containing PISA students noted on Sampling Task 2. Any discrepancies were queried.
- Any school types noted as being excluded were verified as school-level exclusions on the Sampling Task 7b form. Any discrepancies were queried.

#### ***Sampling task 8b: Sampling frame***

- On the spreadsheet for school-level exclusions, the number of schools and the total enrolment figures, as well as the reasons for exclusion, were checked to ensure correspondence with values reported on the Sampling Task 7b form detailing school-level exclusions. It was verified that this list of excluded schools did not have any schools which were excluded for having only one or two PISA-eligible students, as these schools were not to be excluded from the school sampling frame. Checks were done to ensure that excluded schools did not still appear on the other spreadsheet containing the school sampling frame.
- All units on the school sampling frame were confirmed to be those reported on the Sampling Task 2 as sampling frame units. The sampling unit frame number was compared to the corresponding frame for PISA 2012 as well as previous cycles. Differences were queried.
- NPMs were queried about whether or not they had included schools with grades 7 or 8, or in some cases those with grades 10 or higher, which could potentially have PISA-eligible students at the time of assessment even if the school currently did not have any.
- NPMs were queried about whether they had included vocational or apprenticeship schools, schools with only part-time students, international or foreign schools, schools not under the control of the Ministry of Education, or any other irregular schools that could contain PISA-eligible students at the time of the assessment, even if such schools were not usually included in other national surveys.
- The frame was checked for all required variables: a national school identifier with no duplicate values, a variable containing the school enrolment of PISA-eligible students, and all the explicit and implicit stratification variables. Stratification variables were checked to make sure none had missing values and only had levels as noted on Sampling Task 2.
- Any additional school sampling frame variables were assessed for usefulness. In some instances other variables were noted on the school frame that might also have been useful for stratification.
- The frame was checked for schools with only one or two PISA-eligible students. If no schools were found with extremely low counts, but the country's previous sampling frames had some, this was queried.
- The frame was checked for schools with zero enrolment. If there were none, this was assessed for reasonableness. If some existed, it was verified with the NPM that these schools could possibly have PISA-eligible students at the time of the assessment.

#### ***Sampling task 9: Treatment of small schools and the sample allocation by explicit strata***

- All explicit strata had to be accounted for on the form for Sampling Task 9.
- All explicit strata population entries were compared to those determined from the sampling frame.
- All small-school analysis calculations were verified.
- It was verified that separate small-school analyses were done for adjudicated or non-adjudicated oversampled regions (if these were different from explicit strata).

- Country specified sample sizes were monitored, and revised if necessary, to be sure minimum sample sizes were being met.
- The calculations for school allocation were checked to ensure that schools were allocated to explicit strata based on explicit stratum student percentages and not explicit stratum school percentages, that all explicit strata had at least two allocated schools, and that no explicit stratum had only one remaining non-sampled school.
- It was verified that the allocation matched the results of the explicit strata small school analyses, with allowances for random deviations in the numbers of very small, moderately small, and large schools to be sampled in each explicit stratum.
- The percentage of students in the sample for each explicit stratum had to be approximate to the percentage in the population for each stratum (except in the case of oversampling).
- The overall number of schools to be sampled was checked to ensure that at least 150 schools would be sampled.
- The overall number of students to be sampled was checked to ensure that at least 6 300 students would be sampled in CBA countries and 5 250 students would be sampled in PBA countries.
- Previous PISA response rates were reviewed and if deemed necessary, sample size increases were suggested.

#### **Sampling task 10: School sample selection**

- All calculations were verified, including those needed for national study overlap control.
- Particular attention was paid to the required four decimal places for the sampling interval and the generated random number.
- The frame was checked for proper sorting according to the implicit stratification scheme, for enrolment values, and the proper assignment of the measure of size value, especially for very small and moderately small schools. The assignment of replacement schools and PISA identification numbers were checked to ensure that all rules established in the *Sampling Preparation Manual* were adhered to.

#### **Sampling task 11: Reviewing and agreeing to the Sampling Form**

- The form for Sampling Task 11 was prepared as part of the sample selection process. After the international contractor verified that all entries were correct, NPMs had one month to perform the same checks and to agree to the content in this form.

#### **Sampling task 12: School participation and data validity checks**

- Extensive checks were completed on Sampling Task 12 data since it would inform the weighting process. Checks were done to ensure that school participation statuses were valid, student participation statuses had been correctly assigned, and all student sampling data required for weighting were available and correct for all student sampling options. Quality checks also highlighted schools having only one grade with PISA-eligible students, only one gender of PISA-eligible students, or schools which had noticeable differences in enrolled student counts than expected based on sampling frame enrolment information. Such situations were queried.
- Large differences in overall grade and gender distributions compared to unweighted 2012 data were queried.
- Uneven distributions of student birth months were queried when such distributions differed from unweighted 2012 data.
- These data also provided initial unweighted school and student response rates. Any potential response rate issues were discussed with NPMs if it seemed likely that a non-response bias report might be needed.
- Large differences in response rates compared to PISA 2012 were queried.

### **STUDENT SAMPLES**

Student selection procedures in the main study were the same as those used in the field trial. Student sampling was undertaken using the international contractor software, KeyQuest, at the national centres from lists of all PISA-eligible students in each school that had agreed to participate. These lists could have been prepared at national, regional, or local levels as data files, computer-generated listings, or by hand, depending on who had the most accurate information. Since it was important that the student sample be selected from accurate, complete lists, the lists needed to be prepared slightly in advance of the testing period and had to list all PISA-eligible students. It was suggested that the lists be received one to two months before the testing period so that the NPM would have adequate time to select the student samples.



Three countries (Germany, Iceland and Italy) chose student samples that included students aged 15 and/or enrolled in a specific grade (e.g. grade 10). Thus, a larger overall sample, including 15-year-old students and students in the designated grade (who may or may not have been aged 15) was selected. The necessary steps in selecting larger samples are noted where appropriate in the following details:

- Germany supplemented the standard sampling method with an additional sample of grade-eligible students which was selected by first selecting a grade 9 class within PISA-sampled schools that had this grade. In the past, Germany assessed all the class-sampled students. This was not desired for their PISA 2015 national grade 9 sample option. For PISA 2015, to reduce the number of students needing to be assessed for their grade 9 sample from the sampled class, Germany randomly sub-sampled 15 students eligible for the class sample only to participate; the other students eligible only for the class sample were treated as non-respondents. Since non-response in this case was random, these students were accounted for in the grade 9 optional sample through student non-response adjustments.
- Iceland used the standard method of direct student sampling. The sample constituted a de facto grade sample because nearly all of the students in the grade to be sampled were PISA-eligible 15 year olds.
- Italy selected a grade 10 sample by selecting a sample of grade 10 classes. All students from the selected classes were included in the sample.

Four countries (Canada, Denmark, Luxembourg, and Mexico) selected, in addition to PISA students, national-option-eligible-only students to also do the PISA assessments.

### **Preparing a list of age-eligible students**

Each school participating in PISA had to prepare a list of age-eligible students that included all 15 year olds (using the appropriate 12-month age span agreed upon for each participating country) in international grades 7 or higher. In addition, each school drawing an additional grade sample also had to include grade-eligible students that included all PISA-eligible students in the designated grade (e.g. grade 10). In addition, if a country had chosen the international option of the Teacher Questionnaire (see below), eligible teachers were also listed on this form. This form was referred to as a student listing form. The following were considered important:

- Age-eligible students were all students born in 1999 (or the appropriate 12-month age span agreed upon for the participating country). With additional grade samples, including grade-eligible students was also important.
- The list was to include students who might not be tested due to a disability or limited language proficiency.
- Students who could not be tested were to be excluded from the assessment after the student listing form was created and after the student sample was selected. It was stressed to national centres that students were to be excluded after the student sample was drawn, not prior.
- It was suggested that schools retain a copy of the student list in case the NPM had to contact the school with questions.
- Student lists were to be up-to-date close to the time of student sampling rather than a list prepared at the beginning of the school year.

### **Selecting the student sample**

Once NPMs received the list of PISA-eligible students from a school, the student sample was to be selected and the list of selected students returned to the school via a student tracking form. An equal probability sample of PISA students was selected, using systematic sampling, where the lists of students were first sorted by grade and gender. NPMs were required to use KeyQuest, the international contractor sampling software, to select the student samples unless otherwise agreed upon. For PISA 2015, all countries used KeyQuest.

### **Preparing instructions for excluding students**

PISA was a timed assessment administered in the instructional language(s) of each participating country and designed to be as inclusive as possible. For students with limited assessment language(s) experience or with physical, mental, or emotional disabilities who could not participate, PISA developed instructions in cases of doubt about whether a selected student should be assessed. NPMs used the guidelines to develop any additional instructions; school co-ordinators and test administrators needed precise instructions for exclusions. The national operational definitions for within-school exclusions were to be clearly documented and submitted to the international contractor for review before testing.



## Sending the student tracking form to the school co-ordinator and test administrator

The school co-ordinator needed to know which students were sampled in order to notify students, parents, and teachers, and in order to update information and to identify students to be excluded. The student tracking form was therefore sent approximately two weeks before the testing period. It was recommended that a copy of the tracking form be kept at the national centre and the NPM send a copy of the form to the test administrator in case the school copy was misplaced before the assessment day. The test administrator and school co-ordinator manuals (see Chapter 6) both assumed that each would have a copy.

In the interest of ensuring that PISA was as inclusive as possible, student participation and reasons for exclusion were separately coded in the student tracking form. This allowed for special education needs (SEN) students to be included when their needs were not serious enough to be a barrier to their participation. The participation status could therefore detail, for example, that a student participated and was not excluded for special education needs reasons even though the student was noted with a special education need. Any student whose participation status indicated they were excluded for special education needs reasons had to have an SEN code that explained the reason for exclusion. It was important that these criteria were followed strictly for the study to be comparable within and across participating countries. School co-ordinators and test administrators were told to include students when in doubt. The instructions for excluding students are provided in the PISA Technical Standards (Annex F).

## TEACHER SAMPLES

New for PISA 2015, a limited number of countries elected to take an international option in which teachers were sampled in each sampled school. Data from the teacher questionnaire (TQ) was intended to be used to add context to student data from the same school, that is, to describe the learning environment of typical 15-year-old students in the country. Therefore, the TQ focused on that grade level that most 15-year-old students in the country attend, or in other words, the national modal grade for 15-year-old students. If an adjacent grade level was attended by one third or more of 15-year-old students in the country, both grade levels were used as modal grades.

A teacher was defined as “one whose primary or major activity in the school is student instruction, involving the delivery of lessons to students. Teachers may work with students as a whole class in a classroom, in small groups in a resource room or one-to-one inside or outside regular classrooms.”

In order to cover a broader variety of perspectives, and guarantee samples that were large enough, teachers who CAN or WILL be teaching the PISA modal grade in a later year were also considered to belong to the teacher target population. This applied also for teachers who had been teaching the modal grade in the past who were still in the school. Thus, sampling for teachers included ALL teachers that were eligible for teaching the modal grade - whether they were doing so currently, had done so before, or will/could do so in the future.

Teachers were listed and sampled in KeyQuest as either part of Population 4 (science teachers) or Population 5 (non-science teachers). The distinction between Population 4 and Population 5 is determined by the meaning of school science. School science includes all school science courses referring to the domains of physics, chemistry, biology, earth science or geology, space science or astronomy, applied sciences, and technology, either taught in the curriculum as separate science subjects or taught within a single ‘integrated-science’ subject. It does NOT include related subjects such as mathematics, psychology, economics, nor possible earth science topics included in geography courses. Teachers of these subjects were included in the non-science teacher sample.

Ten science teachers were sampled in schools having at least that many listed, or all, if there were not ten. Fifteen non-science teachers were sampled in schools having at least that many listed, or all, if there were not 15. Within each teacher population (science and non-science) an equal probability sample of teachers was selected, using systematic sampling where the lists of teachers were first sorted by grade and gender, where grade had codes indicating whether or not the teacher was currently teaching the modal grade.

## DEFINITION OF SCHOOL

Although the definition of a “school” is difficult, PISA generally aims to sample whole schools as the first stage units of selection, rather than programmes or tracks or shifts within schools, so that the meaning of “between school variance” is more comparable across countries.



There are exceptions to this, such as when school shifts are actually more like separate schools than part of the same overall school. However, in some countries with school shifts, this is not the case, and therefore whole schools are used as the primary sampling unit. Similarly, many countries have schools with different tracks/programmes, but generally it is recommended again that the school as a whole should be used as the primary sampling unit. There are some exceptions, such as the schools being split for sampling in previous PISA cycles (trends would be affected if the same practice was not continued), or if there is a good reason for doing so (such as to improve previously poor response rates, differential sampling of certain tracks or programmes is desired, etc.).

Sampling units to be used on school-level frames were discussed with each country before the field trial. Table 4.3 presents the comments from NPMs, in cases where “school” was not the unit of sampling. Where the Sampling Unit column indicates SFRUNITS, this means that the school was the sampling unit. Where it shows SFRUNITO then something else was used, as described in the comments. Table 4.3 shows the extent to which countries do not select schools in PISA, but rather something else.

**Table 4.3 Sampling frame unit [Part 1/2]**

	Sampling unit school/other	Sampling frame units comment
Albania	School	
Algeria	School	
Argentina	Other	Location of schools
Australia	Other	Schools with more than one campus listed as separate entries
Austria	Other	Either whole schools or programmes within schools
Belgium	Other	French and German speaking communities: a combination of whole schools, or pedagogical-administrative units, which may include different tracks and programmes, and which may also include distinct geographical units. Flanders: implantations, which are tracks/programmes taught on a single address/location (administrative address)
Brazil	School	
Bulgaria	School	
Canada	School	
Chile	School	
B-S-J-G (China)	School	
Colombia	Other	“Sedes,” or physical location
Costa Rica	School	
Croatia	Other	School locations
Cyprus*	School	
Czech Republic	Other	Basic school – whole school special and practical school – whole school gymnasium – pseudo schools according to the length of study (4-year gymnasium and 6- or 8-year gymnasium) upper-secondary vocational – pseudo schools (schools with maturate, schools without maturate)
Denmark	School	
Dominican Republic	School	
Estonia	School	
Finland	School	
France	School	
FYROM	School	
Georgia	School	
Germany	School	Exceptions in SEN schools
Greece	School	
Hong Kong (China)	School	
Hungary	Other	Tracks in parts of schools on different settlements
Iceland	School	
Indonesia	School	
Ireland	School	
Israel	School	
Italy	School	
Japan	Other	Programme
Jordan	School	

**Table 4.3 Sampling frame unit [Part 2/2]**

	Sampling unit school/other	Sampling frame units comment
Kazakhstan	School	
Korea	School	
Kosovo	School	
Latvia	School	
Lebanon	School	
Lithuania	School	
Luxembourg	School	
Macao (China)	School	
Malaysia	School	
Malta	School	
Mexico	School	
Moldova	School	
Montenegro	School	
Netherlands	Other	Locations of (parts of) schools, often parts of a larger managerial unit
New Zealand	School	
Norway	School	
Peru	School	
Poland	School	
Portugal	Other	Cluster of schools; almost all schools are organised in clusters with a unique principal and teachers belonging to each cluster
Puerto Rico (USA) <sup>1</sup>	School	
Qatar	School	
Romania	Other	School programmes
Russian Federation	School	
Scotland	School	
Singapore	School	
Slovak Republic	School	
Slovenia	Other	Study programme within ISCED3 schools and whole ISCED2 schools
Spain	Other	Whole school is the option selected for Spain. Only in the Basque Country (5% of Spanish population) the same school may be divided into three, each one corresponding to each linguistic model (A, B, D) within the region
Sweden	Other	Some schools have been divided horizontally or vertically so that each part has only one principal
Switzerland	School	
Chinese Taipei	School	
Thailand	School	
Trinidad and Tobago	School	
Tunisia	School	
Turkey	School	
United Arab Emirates	Other	Separate curricula and also by gender. Whole schools sometimes.
United Kingdom (excl. Scotland)	School	
United States	School	
Uruguay	School	
Viet Nam	School	

\* See note 1 under Table 4.1.

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.



## **Notes**

1. Students were deemed participants if they responded to at least half of the cognitive items or if they had responded to at least one cognitive item and had completed the background questionnaire (see Annex F).

2. These include schools with multiple languages of mathematics instruction, vocational schools, technical schools, agriculture schools, schools with only part-time students, schools with multiple shifts and so on.

## **References**

**OECD and Westat** (2013), “FT Sampling Guidelines”, report produced by Westat, Core 5 Contractor, for the second meeting of the National Project Managers, March, <https://www.oecd.org/pisa/pisaproducts/PISA2015FT-SamplingGuidelines.pdf>.





5

## Translation and verification of the survey material

<b>Introduction .....</b>	92
<b>Development of source versions .....</b>	92
<b>PISA translation and adaptation guidelines .....</b>	94
<b>Translation training sessions .....</b>	94
<b>Testing languages and translation/adaptation procedures .....</b>	94
<b>Centralised management of changes in trend .....</b>	96
<b>Mode effect study (see Chapter 2).....</b>	96
<b>International verification of the national versions .....</b>	96
<b>Verification of new computer-based test units .....</b>	97



## INTRODUCTION

This chapter explains the procedures used for translation, adaptation and verification for both paper-based (PBA) and computer-based (CBA) materials in PISA 2015.

One of the important aspects of quality assurance in PISA is to ensure that the instruments used in all participating countries to assess students' performance provide reliable and comparable information. In order to achieve this, strict procedures for the localisation (adaptation, translation and validation) of national versions of all survey instrumentation were implemented in PISA 2015 as in all previous rounds.

These procedures included:

- optimising the English source version for translation through translatability assessment
- development of two source versions of the instruments, in English and French (except for the financial literacy and for the operational manuals, provided only in English)
- double-translation design
- preparation of detailed instructions for the localisation of the instruments for the field trial and for their review for the main survey
- preparation of translation/adaptation guidelines
- training of national staff in charge of the translation/adaptation of the instruments
- validation of the translated/adapted national versions: verification by independent verifiers, review by cApStAn staff and the translation referee or the Questionnaires team, countries' post-verification review and "technical" and linguistic final checks.

## DEVELOPMENT OF SOURCE VERSIONS

### Translatability assessment

The translatability assessment was an effort to combine linguists' expertise with that of item developers to bridge the gap between a draft item written in the source language and an actual source version of that item, suitable for translation/adaptation.

While item writers are increasingly aware of localisation issues, they are rarely in a position to identify some of the hurdles translators will be confronted with. In line with the trend to do more upstream work, i.e. work before the start of the actual translation process, a methodology was developed to identify and document potential translation and adaptation difficulties in draft PISA 2015 items before the source versions were finalised. This process, referred to as the translatability assessment, was implemented for the first time in this cycle of PISA.

The translatability assessment consists of submitting draft versions of new items to a pool of experienced linguists covering a broad range of language groups. These individuals were selected among the international verifiers and were trained to use a set of 13 translatability assessment categories to report on potential translation, adaptation and cultural issues they might identify. For both questionnaire items and new science items, the items were submitted in batches. The work was organised so that at least three linguists, from different language groups, would comment on each item.

The approach was for each linguist to first mentally translate each item allocated to him/her. When the item appeared straightforward to translate, the category "straightforward" was selected. When the linguist found an item somewhat difficult to translate/adapt or identified a potential cultural issue, s/he went through the exercise of (i) producing a written translation of that item; (ii) selecting the relevant translatability category; (iii) describing the issue; and (iv) proposing an alternative wording or a translation/adaptation note to circumvent the problem. It should be noted that the translations produced in category (i) were not intended for further use; they were used to help the linguists identify and describe the translation and adaptation hurdles that translators would face if no pre-emptive action were taken.

The feedback from the different linguists was then collated by a senior linguist at cApStAn or, in some cases, by the translation referee: s/he reformulated the comments so that similar issues were processed in a consistent way; selected or rewrote proposals for alternative wording that addressed all the issues identified and drafted translation/adaptation



notes when applicable. When several linguists working in different languages pointed out similar issues in a given item, special attention was given to the wording of that item. The senior linguist produced the *Translatability Report*, which was then sent to the item developers for review. Item developers took this opportunity to eliminate ambiguities, e.g. Anglo-Saxon idiosyncrasies that may be difficult to render in certain languages, double-barrelled questions, cultural issues or unnecessary complexity. Overall, an attempt was made to fine tune the initial version of the items so that it became a more translatable source version.

### **Production of the second source version in French**

Since the inception of the survey, it has been a requirement in the PISA Terms of Reference that the international contractor should produce an international French source version of the data collection instruments. Experience has shown that some issues do not become apparent until there is an attempt to translate the instruments. As in previous PISA survey administrations, the English-to-French translation process proved to be very effective in detecting residual errors overlooked by the test developers, and in anticipating potential problems for translation in other languages. In particular, a number of ambiguities or pitfall expressions could be spotted and avoided from the beginning by slightly modifying both the English and French source versions; the list of aspects requiring national adaptations could be refined; and further translation notes could be added as needed.

The French source version was produced through the double-translation and reconciliation process, followed by a review by a French domain expert for appropriateness of the terminology, and by a native professional French proof-reader for linguistic correctness. In addition, an independent verification of the equivalence between the final English and French versions was performed using the same procedures and verification checklists as for the verification of all other national versions.

Both the translatability assessment and the development of the French source version contributed to providing national project managers (NPMs) with source material that was easier to translate and contained fewer potential translation problems than would have been the case had only one source been developed without a translatability assessment.

### **Double translation from two source languages**

Back translation has long been the most frequently used way to ensure linguistic equivalence of test instruments in international surveys. It requires translating the source version of the test (generally English language) into the national languages, then translating them back to English and comparing them with the source language to identify possible discrepancies. A second approach is a double-translation design (i.e. two independent translations from the source language(s), and reconciliation by a third person). This offers two significant advantages in comparison with the back-translation design:

- Equivalence of the source and target versions is obtained by using three different people (two translators and a reconciler) who all work on both the source and the target versions. On the other hand, in a back-translation design the first translator is the only one to simultaneously use the source and target versions.
- Discrepancies are recorded directly in the target language instead of in the source language, as would be the case in a back-translation design.

Both back-translation and double-translation designs have a potential disadvantage in that the equivalence of the various national versions depends exclusively on their consistency with a single source version (in general, English). In particular, one would wish the highest possible semantic equivalence since the principle is to measure access that students from different countries would have to a same meaning, through written material presented in different languages. Using a single reference language is likely to give undue importance to the formal characteristics of that language. If a single source language is used, its lexical and syntactic features, stylistic conventions and the typical patterns it uses to organise ideas within the sentence will have a greater impact on the target language versions than desirable (Grisay, 2003). The recommended approach in PISA therefore builds on the strengths of the double-translation approach by using double translation from two different source languages.

Resorting to two different languages may, to a certain extent, reduce problems linked to the impact of cultural characteristics of a single source language. Admittedly, both languages used in PISA share an Indo-European origin. However, they do represent relatively different sets of cultural traditions, and are both spoken in several countries with different geographic locations, traditions, social structures and cultures.

The use of two source languages in PISA results in other anticipated advantages such as the following:

- Many translation problems are due to idiosyncrasies: words, idioms, or syntactic structures in one language appear untranslatable into a target language. In many cases, the opportunity to consult the other source version may provide hints at solutions.
- The desirable or acceptable degree of translation freedom is very difficult to determine. A translation that is too faithful to the original version may appear awkward; if it is too free or too literary it is very likely to jeopardise equivalence. Having two source versions in different languages, with clear guidelines on the amount of translation fidelity/freedom, provides national reconcilers with accurate benchmarks in this respect, which neither back translation nor double translation from a single language could provide.

As in previous PISA cycles, the double-translation and reconciliation procedure was a requirement for all national versions of test and questionnaire instruments used in the assessment. It was possible for countries to use the English source version for one of the translations into the national language and the French source version for the other. An efficient alternative method was to perform double translation and reconciliation from one of the source languages, and extensive cross checks against the second source language. Financial Literacy units were double translated from English only, as there was no French source version of these units.

## PISA TRANSLATION AND ADAPTATION GUIDELINES

PISA Translation and Adaptation Guidelines were produced to guide the national teams in the adaptation work of the instruments. The guidelines included:

- Instructions on double or single translation: Double translation (and reconciliation) is required for test and questionnaire materials, but not for manuals, coding guides and other logistic material. In double translation, it is recommended that one independent translator uses the English source version while the second uses the French version. In countries where the National Project Manager (MPM) has difficulty appointing competent translators from French and English, double translation from English or French only is considered acceptable; in such cases it is highly recommended to use the other source version for cross checks during the reconciliation process insofar as possible.
- Instructions on recruitment and training.
- Security requirements.
- References to other documents, including technical guides for translating and reconciling computer-based materials.
- Recommendations to avoid common translation traps.
- Instructions on how to adapt the test material to the national context.
- Instructions on how to translate and adapt questionnaires and manuals to the national context.

In addition to the generic translation and adaptation guidelines, the translators and reconcilers were given item-specific guidelines within the monitoring sheets that accompanied the materials throughout the localisation process. These guidelines provided help for specific translation and adaptation challenges. The item-specific guidelines were produced based on a thorough review first of the English source, then of the comments arising from the translatability assessment and then of those arising from the production of the French source version.

## TRANSLATION TRAINING SESSIONS

National project managers received sample materials to use when recruiting national translators and training them at the national level. The NPM meeting held in March 2013 in Bangkok included sessions on the field trial translation/adaptation activities in which recommended translation procedures, PISA Translation and Adaptation Guidelines, and the verification process were presented in detail separately for each component of the survey (questionnaires, collaborative problem-solving units, new scientific literacy units, trend units).

## TESTING LANGUAGES AND TRANSLATION/ADAPTATION PROCEDURES

National project managers had to identify the testing languages according to instructions given in the School Sampling Preparation Manual and to record them in a sampling form for agreement.

Prior to the field trial, national project managers had to fill in a translation plan describing the procedures used to develop their national versions and the different processes used for translator/reconciler recruitment and training. Information



about a possible national expert committee was also sought. This translation plan was reviewed by the translation referee for discussion/approval.

Figure 5.1 summarises the field trial translation procedures for tests and questionnaires, as described in the confirmed translation plans. The figures in the table include minority language versions that represented less than 10% of the target population and were not verified internationally.

■ Figure 5.1 ■

#### Translation procedures reported by national centres in the translation plan

	Tests	Questionnaires
Double translation from English and French source versions	20	15
Double translation from English source version with cross checks against the French source version	8*	20*
Double translation from French source version with cross checks against the English source version	1	1
Double translation from English source version only	23	25
Double translation from French source only	1	1
Adaptations in one of the source versions	26	26
Adaptations made in a borrowed verified version or "base" version	34	24

\* For the Catalan, Galician (questionnaires only) and Basque versions, the cross checks were made against the verified Spanish version of Spain.

Note: The totals do not match between tests and questionnaires, because in the case of the German version, the procedure used was different for new science and collaborative problem solving units.

The lower number of questionnaire versions adapted from a verified or base version versus the same number for tests is largely explained by the fact that a Spanish base version of the tests was produced, as described below, but there was no Spanish base version of the questionnaires. Therefore, countries that could adapt the Spanish base version for test units were responsible for translating the questionnaires themselves. Regarding the lower number of questionnaire versions translated from both English and French compared to tests, this is a known trend over all PISA cycles. However, this decrease was amplified for PISA 2015 because the French source version was only made available as a word document; the "online" version was available in English only. Countries therefore preferred to use French for cross checks only.

As in PISA 2012, when mathematics was the major domain, there is a "domain effect" in the translation procedures compared to PISA 2009, when reading literacy was the major domain. Some countries (e.g. Germany and Norway) that used double translation from both English and French sources in 2009 chose double translation from the English source with cross checks against the French source version in 2015 because they could not find translators from French with good experience in the scientific literacy domain.

Countries sharing a testing language were strongly encouraged to develop a common version in which national adaptations would be inserted or, in the case of minority languages, to borrow an existing verified version. It has been found in previous survey administrations that high-quality translations and high levels of equivalence in the functioning of items were achieved in countries that shared a common language of instruction and could develop their national versions by introducing a limited number of national adaptations in a common version. Additionally, a common version for different countries sharing the same testing language implies that all students instructed in a given language receive booklets that are as similar as possible, which reduces cross-country differences due to translation effects.

Co-operation between countries sharing a same language was therefore fostered and facilitated: workable models were designed so that verified versions from one country could be adapted by a second country.

- As in previous cycles, the model followed by German-speaking countries was (again) highly efficient: the German version of each of the components of the assessment material was double translated and reconciled by one of the countries, then verified, and adapted by the other countries who administered that component. The adapted versions were then verified.
- A Spanish base version of the new test materials was produced by an independent contractor and shared by seven Spanish-speaking countries (Chile, Colombia, Costa Rica, Dominican Republic, Peru, Spain and Uruguay) – only Mexico opted for an independent translation; Argentina also tested in Spanish but was a paper-based country so did not use the new test materials.

Translation of coding guides for open-ended items was not included in the translation plan because, for PISA 2015, the recommended procedure was to single-translate from one source version with cross checks against the other. Some countries produced translated coding guides in one national language only (Spain), while some used the English source (Sweden) or French source (Tunisia) without translation.

## CENTRALISED MANAGEMENT OF CHANGES IN TREND

In PISA 2015, a centralised management approach for trend content was implemented for both test and questionnaire materials. The cornerstone of this approach is that all changes to trend content requested by countries went through a strict negotiation process; approved changes were then implemented centrally so that countries did not have editing rights at any stage of the process. This approach prevents unnecessary, undocumented or unverified changes in the trend materials, and thus will allow both more reliable comparability across cycles, and a detailed record of all changes made in trend materials.

## MODE EFFECT STUDY (SEE CHAPTER 2)

To enable study of mode effects, all computer-based assessment (CBA) countries (with the exception of Austria, due to a delayed testing window) administered their trend units in both computer-based (CBA) and paper-based (PBA) mode. As part of the centralised trend management process, all changes made to the CBA version of a trend unit were also reflected in the PBA version of the same unit, so that consistency between the same unit administered in two different delivery modes could be maintained.

## INTERNATIONAL VERIFICATION OF THE NATIONAL VERSIONS

As in previous PISA survey administrations, one of the most important quality control procedures implemented to ensure high-quality standards in the translated assessment materials for PISA 2015 was to have an independent team of expert verifiers, appointed and trained by the international contractors, verify each national version against the English and/or French source versions.

International verification was carried out for all national versions in languages used in schools attended by more than 10% of the country's target population.

The main criteria used to recruit verifiers of the various national versions were that they had:

- native command of the target language
- professional experience as translators from English or French or from both English and French into their target language
- as far as possible, sufficient command of the second source language (either English or French) to be able to use it for cross checks in the verification of the material. Note that not all verifiers are proficient in French, but this is mitigated by the fact that the cApStAn reviewer and the translation referee have command of French
- as far as possible, familiarity with the main domain assessed, in this case, scientific literacy
- a good level of computer literacy and experience with computer-aided translation tools (CAT tools)
- as far as possible, experience as teachers and/or higher education degrees in psychology, sociology or education.

A verifier training seminar was held prior to the verification of the field trial materials. For those who could not attend the training seminar, webinars were organised. The training sessions focused on:

- presenting verifiers with PISA objectives and structure
- familiarising them with the material to be verified, the verification procedures, and the software tools to be used (in particular, the open language tool (OLT) software used for computer-based materials)
- reviewing and extensively discussing the translation guidelines and the verification checklist
- conducting hands-on exercises on specially “doctored” target versions in which typical errors (linguistic issues, adaptation issues, or errors related to guidelines not being followed) had been planted
- arranging schedules and dispatch logistics
- security requirements.



Verification procedures have been continually improved throughout each PISA round, based on the experience and learning from previous rounds. In PISA 2015, the change from paper-based delivery mode into computer-based delivery mode also brought changes in the procedures. In the following subsections we review the procedures implemented in PISA 2015 for the different components subject to verification.

## VERIFICATION OF NEW COMPUTER-BASED TEST UNITS

Fifteen of the countries in PISA 2015 participated in the paper-based assessment (PBA), while the rest participated in the computer-based assessment. This was a significant change from PISA 2012 where the main delivery mode was still paper-based.

Computer-based units were translated and verified using the open language tool (OLT) software on XLIFF (tagged XML Localisation Interchange File Format) files which were exchanged, previewed and archived on the PISA portal, a web-based platform that allows the files to travel through a predefined workflow.

To perform the verification task, the verifiers were instructed to verify the text segments one by one, comparing the target version appearing on the right side of the OLT interface to the source version appearing on the left side, while consulting previews on the portal and the test adaptation spreadsheet (TAS) to see item-specific guidelines and comments from the national centres. They made corrections as needed, documenting their interventions in the test adaptation spreadsheet, including selection of the appropriate intervention category using a drop-down menu.

Once a domain was verified, reviewed and “finalised” on the portal, the translation referee was able to download the test adaptation spreadsheet annotated by the verifier. The referee would then go through each verifier comment, and label as “requires follow-up” any crucial issues that could potentially affect equivalence or item functioning. Changes labelled as “requires follow-up” were negotiated between the referee and the national centre. The national centre then uploaded revised XLIFF files on the portal for final check. The final check reviewer checked the correct implementation of any changes “requiring follow-up” and either released the files for layout check and national version construction by the international contractors or released them back to the national centre for additional corrections.

Since the PISA 2003 main survey, the central element and repository of the entire translation, adaptation and verification procedure for test units has been the test adaptation spreadsheet. Figure 5.2 shows a sample test adaptation spreadsheet from the PISA 2015 field trial. The spreadsheet functions as:

- an aid to translators, reconcilers, and verifiers through the increasing use of item-specific translation/adaptation guidelines
- a centralised record of national adaptations, of verifier corrections and suggestions
- a way of conducting discussions between the national centre and the translation referee
- a record of the implementation status of “requires follow-up” in test units
- a tool permitting quantitative analysis of verification outcomes.

■ Figure 5.2 ■

### Sample of a test adaptation spreadsheet (TAS) from the PISA 2015 field trial

ENGLISH SOURCE VERSION	ITEM-SPECIFIC TRANSLATION / ADAPTATION GUIDELINE	COUNTRY COMMENT (ADAPTATION, DOUBTS)	VERIFIER INTERVENTION	VERIFIER COMMENT	CONSORTIUM REFEREE COMMENT	CORRECTION STATUS	COUNTRY POST-VERIFICATION COMMENT	FINAL CHECK
Refer to “...” on the right. Type your answer to the question.	Recurring instructions	OK	Inconsistency	1st instruction harmonised with SC645, seg 4	Please make sure to keep the verifier correction		OK	
			Register/Wording	“Stress builds up...” translated as “Stress creates...” Verifier thinks translation in the meaning of accumulating/increasing is more appropriate. Changed by ver.	Please keep the verifier correction	REQUIRES FOLLOW-UP	OK	OK

## Verification of homolinguual versions

Whenever a country adapted their national version from the English or French source, the Spanish base version or a same-language verified version of another country, the resulting national version was verified using a special procedure for these so-called homolinguual versions. There were in total 34 national versions that were verified using this process.

The essential difference between the “full” verification of translated national versions and the “focused” verification of homolinguistic versions is that in the latter, the verification concentrates on the changes made by the country versus the base, source or borrowed version. Automatically-created difference reports were used to identify all such changes in a reliable way.

### **Verification of paper-based test units and booklet shell**

Since no new paper-based units were developed for PISA 2015, PBA countries that had participated in cycles 2003, 2006, 2009 and 2012 did not have anything new to translate or adapt. For them, the units only went through the centralised change-management process where the country had the opportunity to request corrections to errors, and these – when accepted by the translation referee – were then implemented centrally by the verifiers.

Paper-based countries that were new in PISA 2015 or that had not participated in one or more of the relevant cycles had to translate or adapt any units they had not administered before. These were verified following the same process as described above for computer-based materials. The only essential difference was that the verifiers implemented the changes in the MS Word files using the “track changes” functionality, rather than in the online system. The test adaptation spreadsheet was used the same way as in the computer-based verification.

### **Verification of questionnaires**

Questionnaires were submitted for verification together with an agreed questionnaire adaptation spreadsheet (QAS). The first purpose of the questionnaire adaptation spreadsheet was to document all content-related or ‘structural’ deviations from the international reference versions. Such national adaptations were subject to clearance by the questionnaire team before the material was submitted for verification. Subsequently, the spreadsheet served the same objectives and followed the same logic as the test adaptation spreadsheet for test units (see above). Figure 5.3 shows a sample questionnaire adaptation spreadsheet from the PISA 2015 field trial.

■ Figure 5.3 ■

**Sample of a questionnaire adaptation spreadsheet (QAS) from the PISA 2015 field trial**

National Centre to complete			Questionnaire Team to complete			Verifier to complete		
8a English translation of the national version	8b Proposed target version in national language	9 Justification for proposed changes; national centre comments	10 Questionnaire team Comments	11 Recode suggestion or other	12 Agreement status	13 Verifier intervention category	14 Verifier comments	15 Verifier target version
Number of lessons per week in Slovak language	Počet vyučovacích hodín slovenského jazyka a literatúry týždenne				AGREED	OK	Complete translation in 8a should be: “Number of lessons per week in Slovak language and literature,” which is the whole correct name of the subject. Considered OK by ver.	
Number of lessons per week in mathematics	Počet vyučovacích hodín matematiky týždenne				AGREED	OK		
Number of lessons per week in school science	Počet vyučovacích hodín prírodovednych predmetov týždenne				AGREED	OK	Adapted as: Number of lessons per week in science subjects. Considered OK by ver.	

The purpose of the verifiers’ brief was to check whether or not target questionnaires are linguistically correct and faithful to either the source version (when no adaptation is made) or the approved English translation of the national version (when an adaptation is made). In light of this, verifiers were instructed:

- to check whether the back translation of the agreed adaptation was faithful
- to check whether the agreed adaptation was correctly reflected in the questionnaire
- to check the questionnaires for undocumented adaptations (deviations from the source not listed in the questionnaire adaptation spreadsheet) and report them
- to check linguistic correctness (grammar, spelling, etc.) of the entire target version.

For the paper-based questionnaires (Student and School questionnaires for countries administering paper-based assessment, Parent Questionnaire for all countries taking this option), verifier interventions were entered in the questionnaires using the track changes mode, while verifier comments were entered in the verifier column of the questionnaire adaptation spreadsheet.



For computer-based questionnaires administered on the questionnaire authoring tool (QAT) platform, the verifier interventions were inserted in the spreadsheet in a separate “Verified target version” column, in addition to documenting the rationale for the change. The verifiers did not have editing access to the platform. If the change was agreed, the country implemented it in the spreadsheet. In paper-based questionnaires, the verifier introduced the changes in the Word files using the track changes functionality, and documented the changes in the spreadsheet.

As for test units, any more significant changes were labelled as “requires follow-up” by the questionnaire team, and after negotiation with the country teams, their correct implementation was checked by verifiers during final check.

There were no special “homolinguial” procedures for the verification of questionnaires since differences in education systems mean that these are very extensively adapted even when sharing a common language. Nevertheless, English and French versions benefited from a co-ordination process similar to the one implemented for test materials. A list of “tips” for verification of questionnaires, including spelling, possibly recurring adaptation issues, and especially errata (errors identified in the source version after release to the countries) and “quasi-errata” (suggestions for improving the source) was maintained, built up, and used in each successive verification.

As in PISA 2012, there was also an increased effort to harmonise the verification feedback for different language versions of questionnaires used in the same country (e.g. German, French and Italian for Switzerland, or the five language versions for Spain). Such versions are by necessity entrusted to different verifiers, but as frequently as possible, cApStAn’s verification reviewers made a point of reviewing and delivering such versions together, striving to harmonise verification interventions on adaptation issues common to the different language versions.

### **Verification of coding guides**

In PISA 2015, the coding guides were verified separately from the test items, and at a later time. This was necessary since a large number of additions and improvements were made to the master versions after the coder training meetings, long after preliminary versions had been made available to countries. As in PISA 2012 and contrary to cycles before that, the scoring sections were not made available for translation at the time of the unit dispatch. There was one coding guide per trend domain (mathematics, science and reading). For CBA countries, there was, in addition, one coding guide for new science units, and for those countries that opted for financial literacy, there was a separate coding guide for this domain.

The overall procedure was the same as for paper-based test units: verifier corrections were made in track changes in the MS Word files, and documented in the monitoring sheets in Excel format. For countries that had participated in previous cycles, trend coding guides underwent a similar controlled change request process as the test units.

### **Main survey verification**

In previous cycles, the instruments were revised to some extent between the field trial and main survey and were then re-verified in this revised form before the main survey. In PISA 2015, no changes were made in the master versions after the field trial (apart from entire units or items being dropped), and verification consisted of verifying changes that countries requested to their FT instruments, for example based on poor performance or differential item functioning in the FT, or the detection of residual “outright errors” (the latter, in particular, for questionnaires). This process was similar to the centralised change management used to control changes in trend: countries requested changes, and the verifiers implemented centrally those changes that were approved by the translation referee. The countries did not have editing access to their units or questionnaires at this stage.

### **Quantitative analyses of verification outcomes**

In PISA 2015, the instruments used to document the verification were designed to generate statistics, thus providing some quantitative data on the frequency of different types of issues identified. The verification statistics by item and by unit yielded information on translation and adaptation difficulties encountered for specific items in specific languages or groups of languages. This type of information, when gathered during the field trial gives valuable information on how to avoid such problems in further survey administrations.

This information also makes it possible to detect whether there are items that elicited many verifier interventions in many language groups. When this occurs, item developers would be prompted to re-examine the item’s reliability or relevance. Similarly, observing the number of adaptations that the countries proposed for some items may give the item developers additional insight into how difficult it is for some countries to make such items suitable for their students. While such adaptations may be discussed with the international contractors, it remains likely that extensively adapted items will eventually differ from the source version (e.g. in terms of reading difficulty).





---

**6**

# Field operations

<b>Overview of roles and responsibilities .....</b>	102
<b>The selection of the school sample .....</b>	104
<b>Preparation of test booklets, questionnaires, and manuals .....</b>	104
<b>The selection of the student sample.....</b>	105
<b>Packaging and shipping materials.....</b>	106
<b>Test administration .....</b>	107
<b>Receipt of materials at the national centre after testing .....</b>	108
<b>Main survey review.....</b>	108

## OVERVIEW OF ROLES AND RESPONSIBILITIES

PISA was co-ordinated in each country<sup>1</sup> by a National Project Manager (NPM) who implemented the procedures specified by the international contractors responsible for PISA implementation. Each NPM typically had several assistants working from a base location that is referred to throughout this report as a National Centre. For the school-level operations, the NPM co-ordinated activities with school-level staff, referred to in PISA as School Co-ordinators.<sup>2</sup> Trained Test Administrators administered the PISA assessment in schools.

### National Project Managers

NPMs were responsible for implementing the project within their own country. They:

- attended NPM meetings and received training in all aspects of PISA operational procedures
- negotiated nationally-specific aspects of the implementation of PISA with the international contractors, such as national and international options, oversampling for regional comparisons, additional analyses and reporting (e.g. by language group)
- established procedures for maintaining the security and confidentiality of materials during all phases of the assessment implementation
- determined the general suitability of using school computers to conduct the computer-based assessment (CBA countries only)
- prepared a series of sampling forms documenting sampling-related aspects of the national educational structure
- prepared the school sampling frame and submitted this to the international contractor for the selection of the school sample
- organised for the preparation of national versions of the test instruments, questionnaires, school-level materials (manuals, scripts and forms), and coding guides
- identified School Co-ordinators from each of the sampled schools (nominated by the school principal or a volunteer from the school staff) and worked with them on school preparation activities
- used software to select the student sample from the lists of eligible students provided by the School Co-ordinators
- used software to select the teacher sample from the lists of eligible teachers provided by the School Co-ordinators, if applicable
- recruited and trained Test Administrators according to the PISA 2015 Technical Standards: Standards 9.1, 9.2, 9.3 and 9.4 to administer the assessments within schools (see Annex F)
- nominated suitable persons to work on behalf of the international contractors as external PISA Quality Monitors (PQMs) to observe the assessment administration in a selection of schools (main survey only)
- monitored the completion of School Questionnaires
- monitored the completion of Teacher Questionnaires (if applicable)
- monitored the completion of Parent Questionnaires (if applicable)
- recruited and trained coders to code the open-ended test items and the occupational data on questionnaires
- arranged for the data entry of the test responses, Student Questionnaire responses, and School Questionnaire responses completed on hard copy (paper-based assessment (PBA) countries)
- submitted the national database to the international contractors
- arranged for the transmission of School Questionnaire and Teacher Questionnaire (if applicable) and responses completed online
- arranged for the coding, data management, and reporting on the Parent Questionnaire (if applicable) or other options (if applicable)
- submitted a written review (Main Survey Report) of PISA implementation activities following the assessment.

A National Project Manager's Manual provided detailed information about the duties and responsibilities of the NPM. Supplementary manuals, with detailed information about particular aspects of the project, were also provided and are described in the relevant chapters.



## School Co-ordinators

School Co-ordinators co-ordinated school-related activities with the National Centre and the Test Administrators. A School Co-ordinator's Manual, prepared by the international contractors, described in detail the activities and responsibilities of the School Co-ordinator.

The School Co-ordinator:

- established the assessment date and time, in consultation with the NPM
- ran a systems diagnostic tool provided by the international contractors to determine if school computers were suitable for the assessment (CBA countries)
- prepared the student list with the names of all eligible students in the school and sent it to the National Centre so that the NPM could select the student sample using KeyQuest
- prepared the teacher list with the names of all eligible teachers in the school and sent it to the National Centre so that the NPM could select the teacher sample using KeyQuest (if applicable)
- received the list of sampled students from the NPM on the Student Tracking Form (a form designed to record sampled students with their background data) and updated it if necessary (e.g. identifying students with disabilities or limited assessment language proficiency who could not take the assessment according to criteria established by the international contractors and the PISA Technical Standards)
- received the list of sampled teachers on the Teacher Tracking Form from the NPM (if applicable) and updated it (e.g. identifying teachers who refused to complete the questionnaire, no longer taught at the school, or were otherwise ineligible)
- received, distributed, and collected the School Questionnaire (if on hard copy) or monitored the completion of the School Questionnaire if completed online
- distributed instructions for completing the Teacher Questionnaire online and monitored the completion online (if applicable)
- received and distributed the Parent Questionnaire if applicable (generally, the Test Administrator distributed the Parent Questionnaire to students on the assessment day to give it to their parents to complete, or the School Co-ordinator sent the questionnaire to the parents 1-2 weeks before the assessment and requested that students return it to the School Co-ordinator before the assessment date)
- informed school staff, students, and parents of the nature of the assessment and the assessment date by sending a letter or organising a meeting
- secured parental permission, if required by the school or education system
- informed the NPM and Test Administrator of any assessment date or time changes
- arranged for technical support if administering the assessment on computers
- assisted the Test Administrator with room arrangements for the assessment day.

On the assessment day, the School Co-ordinator was expected to ensure that the sampled students attended the assessment session(s). If necessary, the School Co-ordinator also made arrangements for a follow-up session and ensured that absent students attended the follow-up session.

## Test Administrators

The Test Administrators were primarily responsible for administering PISA fairly, impartially, and uniformly, in accordance with international standards and PISA procedures. To maintain objectivity, a Test Administrator could not be the science, reading, or mathematics teacher of the students being assessed, and it was preferred that they not be a staff member at any participating school (see Standard 8.2 in Annex F). Prior to the test date, Test Administrators were trained by National Centres. Training included a thorough review of the Test Administrator's Manual and the Student Delivery System Manual (CBA countries). Additional responsibilities included:

- ensuring receipt of the testing materials from the NPM and maintaining their security
- contacting the School Co-ordinator one to two weeks prior to the test to confirm plans
- completing final arrangements on the test day

- reviewing and updating the Student Tracking Form
- completing the Session Attendance Form (a form designed to record students' attendance and instruments allocation)
- completing the Session Report Form (a form designed to summarise session times, any disturbance to the session, etc.)
- ensuring that the number of test booklets and questionnaires collected from students tallied with the number sent to the school (PBA countries) or ensuring that the number of USB sticks used for the assessment were accounted for (CBA countries)
- obtaining the School Questionnaire from the School Co-ordinator (PBA countries)
- obtaining Parent Questionnaires and Teacher Questionnaires (if applicable)
- conducting a follow-up session, if needed, in consultation with the School Co-ordinator
- sending the School Questionnaire, Student Questionnaires, Parent and Teacher Questionnaires (if applicable), and all test materials (both completed and not completed) to the National Centre after the testing.

## THE SELECTION OF THE SCHOOL SAMPLE

NPMs used the detailed instructions in the School Sampling Preparation Manual to document their school sampling plan and to prepare their school sampling frame.

The national target population was defined, school- and student-level exclusions were identified, and aspects such as the extent of small schools (a small school is defined as any school whose approximate enrolment falls below the target cluster size) and the homogeneity of students within schools were considered in the preparation of the school sampling plan.

For all but one country, the sampling frame was submitted to the international contractor, who selected the school sample. Having the international contractor select the school sample minimised the potential for errors in the sampling process and ensured uniformity in the outputs for more efficient data processing later (student sampling, data analysis). It also relieved the burden of this task from National Centres. NPMs worked closely with the international contractor throughout the process of preparing the sampling documentation, ensuring that all nationally-specific considerations related to sampling were thoroughly documented and incorporated into the school sampling plan.

## PREPARATION OF TEST BOOKLETS, QUESTIONNAIRES, AND MANUALS

As explained in Chapter 2, the mode study design for the PISA 2015 field trial required all countries to test using the 18 paper-and-pencil forms that included the trend items for reading, mathematics and scientific literacy and, where applicable, the two Financial Literacy booklets. As part of the 2015 quality control process, the contractors assumed responsibility for preparing national versions of the paper-based trend clusters by extracting clusters from existing booklets in the PISA archives and formatting them for the 2015 cycle. Those countries who were new to PISA in 2015, or who were missing units from previous cycles in which they had not participated, translated those materials following the standard PISA translation/reconciliation process. All countries updated and translated the common booklet parts, which were revised for PISA 2015 and included the cover, general instructions, formula sheet for Mathematics, and the acknowledgements page.

Once the clusters and common booklet parts were finalised and approved by the National Centres, the field trial booklets were assembled by the contractors and shared with countries for final review and signoff. Following that approval process, print-ready files were provided to National Centres.

For the main survey, only those countries not testing on computer prepared paper booklets. National Centres were asked to document any errors in the field trial versions of their booklets that required correction. Those requests were reviewed by the contractors and, where appropriate, revisions were made. As was the case in the field trial, booklets were provided to National Centres for final review and sign off, and print-ready files were then provided.

The computer-based version of the tests included both trend and new items. In preparation for the field trial, where countries had existing translations of trend items, the contractors copied those materials into the computer format and then provided those materials for review and revisions. Countries were asked to document any linguistic or layout issues in their computer-based materials and any corrections were implemented by the international contractors. All new computer-based items were translated by national teams following the translation and reconciliation processes defined in the PISA standards (see Chapter 5 for detailed information).



Once all field trial instruments were approved by National Centres, including both the test items and background questionnaires, files were locked and the national Student Delivery Systems (SDS) was prepared. Please see Chapter 2 for more information about SDS preparation and testing.

In preparation for the main survey, computer-based countries were asked to review their items based on the field trial data and identify any serious errors in need of correction. The contractors worked with National Centres to resolve any remaining linguistic or layout issues and prepared the national instruments for the main survey.

In addition to the standard Student Questionnaire, the Information and Communication Technology Familiarity questionnaire and/or the Educational Career Questionnaire were administered, depending on which options were chosen by the individual country. Forty-seven countries administered the Information and Communication Technology Familiarity Questionnaire and 22 countries administered the Educational Career Questionnaire. The standard Student Questionnaire had to be presented first in the questionnaire booklet.

Two PBA countries and 16 CBA countries also administered the optional Parent Questionnaire. Nineteen CBA countries administered the optional Teacher Questionnaire, which was only available on the computer. All countries administered the obligatory School Questionnaire, with all PBA countries doing it on paper and all CBA countries doing it on the computer.

As with the test material, source versions of the questionnaire instruments in both French and English were provided to NPMs for translation into the languages of the test.

NPMs were permitted to add a maximum of five questions of national interest as national add-ons to the questionnaires. Proposals and text for these were submitted to the international contractor for approval as part of the process of reviewing adaptations to the questionnaires. It was required that the additional material should be placed at the end of the international modules. Following approval of adaptations, the material was verified by the international contractor. NPMs implemented feedback from verification in the assembly of their questionnaires. For paper-based countries, PDFs were finalised by the NPM and uploaded to the PISA Portal. For the computer-based instruments, contractors cross-checked and implemented final versions in all languages on the computer-based platform. More information is given in Chapter 17 of this report.

The School Co-ordinator's Manual, Test Administrator's Manual, and script(s) used to administer the various sessions (these include the Test Administrator's Script, the *Une Heure* (UH) Script [if applicable], and the Financial Literacy Script [if applicable]) were also required to be translated into the national test language(s). Only English source versions of the manuals and scripts were provided by the international contractors. NPMs were required to make adaptations to the manuals and script(s) using the New Comment and Track Changes functions in Microsoft Word. Alternatively, NPMs could submit a Materials Adaptation Spreadsheet (MAS) documenting all proposed national adaptations to the manuals and script(s) to the international contractor for approval. Only a few countries used the MAS. Following approval of the adaptations, the manuals and scripts were translated in the national test language(s).

In countries with multiple assessment languages, the assessment instruments, manuals, and script(s) needed to be translated into each assessment language. For a small number of countries, where Test Administrators were bilingual in the assessment language and the national language, it was not required for the manuals to be translated into both languages. However, in these cases, it was still a requirement that the script(s) was/were translated into the language of the test.

Various checking procedures were employed to review how closely national translations of the school-level materials (manuals and scripts) adhered to the Technical Standards. Key elements of the adapted national language versions were reviewed in approximately 10% of countries. No significant deviations were noted that might affect data validity and reliability. During the main survey, PISA Quality Monitors (PQMs) in all countries were asked to compare the adapted English source versions with the national translations. PQMs questioned the translations in about 5% of countries and these translations were then reviewed. Again, no significant deviations were noted that might impact data quality.

## THE SELECTION OF THE STUDENT SAMPLE

Following the selection of the school sample by the international contractor, the list of sampled schools was returned to National Centres. NPMs then contacted these schools and requested a list of all PISA-eligible students from each school. This was used by NPMs to select the student sample.

NPMs were required to select the student sample using KeyQuest, the PISA student sampling software prepared by the international contractor. KeyQuest generated the list of sampled students for each school, known as the Student Tracking Form, and the Session Attendance Form that served as the central administration documents for the study and linked students, test booklets, and student questionnaires.

## PACKAGING AND SHIPPING MATERIALS

The following key documents and items needed to be sent either to the Test Administrator or to the school:

- test booklets and Student Questionnaires for the number of students sampled plus extra unassigned booklets and questionnaires (PBA countries)
- Student Tracking Form
- Session Attendance Forms, which were specific to PBA or CBA countries (a separate Session Attendance Form was used for Financial Literacy sessions in countries that selected this international option)
- Session Report Form(s)
- test delivery USB sticks (CBA countries)
- Student Logon Forms (CBA countries)
- results from the school's computer system diagnostic report to determine the suitability of running the computer-based assessment from a USB stick (CBA countries)
- Materials Reception Form
- Materials Return Form
- additional materials, e.g. pens and calculators, per local circumstances<sup>3</sup>.

For PBA countries, one of the 18 separate test booklets in the field trial and one of the 30 separate test booklets in the Main Survey was pre-allocated to each student by the KeyQuest software from a random starting point in each school. KeyQuest was then used to generate the school's Session Attendance Forms, which contained the number of the allocated booklet alongside each sampled student's name. This information was used by the Test Administrators when distributing the booklets to students.

For CBA countries, due to the mode effect study, during the field trial, both paper booklets and computer-based forms were assigned automatically by the KeyQuest software. For the Main Survey, there was no paper-based mode. Computer-based forms were assigned automatically by KeyQuest based on the integrated design.

### Field operations procedures specific to paper-based assessment countries

It was recommended that National Centres print removable labels, each with a student identification number and his or her specific test booklet number, as well as the student's name, if this was an acceptable procedure within the country. Two or three copies of each student's label could be printed and used to identify the test booklet and the questionnaire. After the assessment, labels were removed to help ensure the confidentiality of students' responses.

NPMs were allowed some flexibility in how the materials were packaged and distributed, depending on national circumstances. In most countries, materials were shipped directly to the independent Test Administrator rather than to the school. It was specified, however, that the test booklets for a school be packaged so that they remained secure, possibly by wrapping them in clear plastic and then heat-sealing the package, or by sealing each booklet in a labelled envelope. Most countries bundled booklets specific to a school, and the Test Administrator applied the removable student labels prior to the test date. Procedures for preparing test booklets and student questionnaires were described in the Test Administrator's Manual.

### Field operations procedures specific to computer-based assessment countries

It was highly recommended that Test Administrators test the USB sticks prior to the test day to detect any defective USB sticks. Directions for testing the USB sticks were provided in the Student Delivery System Manual.

Test Administrators prepared the Student Logon Forms by ordering them in the order that the students appeared on the Session Attendance Form, numbering the Student Logon Forms, and then crosschecking that the password listed on the Session Attendance Form matched the password listed for that student on the Student Logon Form.



NPMs were allowed some flexibility in how the materials were packaged and distributed, depending on national circumstances. In most countries, materials were shipped directly to the independent Test Administrator rather than to the school.

## TEST ADMINISTRATION

After arriving at the school on assessment day, Test Administrators were required to review the Student Tracking Form with the School Co-ordinator and update the form as necessary. The Session Attendance Forms also were updated as necessary. Once the forms were updated, the Test Administrator set up the room and materials for the assessment session following the steps as described in the Test Administrator's Manual:

### Steps for setting up CBA test administration

- allocate a work space and computer to each participating student.
- set up computers for each student expected to be tested.
- distribute Student Logon Forms to students, ensuring that each student receives only the logon form assigned to that student on the Session Attendance Form.
- set aside the materials for students who had any non-participant codes recorded on the Student Tracking Form or did not attend the assessment session from the very beginning.

### Steps for setting up PBA test administration

- allocate a work space to each participating student.
- distribute test booklets (and later Student Questionnaires) to students, ensuring that each student receives only the test booklet assigned on the Session Attendance Form.
- write the testing date on a board visible to all students.
- ask the students to write the testing date on their test booklet covers in the required format DD/MM/YYYY at the beginning of the session.
- set aside the materials for students who had any non-participant codes recorded on the Student Tracking Form or did not attend the assessment session from the very beginning.

### Administering and monitoring the test

To obtain comparable and reliable data, Test Administrators were required to follow the timing of the paper-based assessment strictly, especially the administration of the test sessions (2 sessions of exactly 1 hour) shown in Figure 6.1 below. The timings were the same for CBA test sessions, with additional time added if a country was administering one of the optional questionnaires. Although CBA test sessions were timed by the test system programme, Test Administrators were still required to enforce the timing and not move students forward prematurely.

■ Figure 6.1 ■  
**Timing of paper-based assessment**

Activity	Timing
Distributing materials and reading the General Directions	15 minutes (approximately)
First 60 minutes of test	60 minutes (exactly)
Short break	Generally, no more than 5 minutes
Second 60 minutes of test	60 minutes (exactly)
Break	15 minutes*
Student Questionnaire	35 minutes (approximately)
Collecting the materials and ending the session	15 minutes (approximately)
<b>Total</b>	<b>Student Time: 3 hours 30 minutes (approximately)</b>

\* The amount of break time before beginning the Student Questionnaire is not absolute. The recommended amount of time is 15 to 30 minutes, but the time can be longer or shorter depending on the discretion of the National Centre, and school circumstances.

NPMs were allowed to adapt the length of the short break after the first hour of testing. Most countries allowed only the recommended 5-minute break. In a few cases, countries did not offer a break between test sections as they felt this would be too disruptive.

No changes to the timing of the test sessions were allowed. Adaptation to the timing of the Student Questionnaire session (both CBA and PBA) was more flexible to maximise the contextual data obtained from students.

The test scripts for both CBA and PBA countries had to be read to the students word-for-word to maintain standardised assessment procedures across all participating countries. For PBA countries, the Test Administrators were required to read the extensive practice exercises and other key instructions to the students. Therefore, if students arrived after these instructions were read, the student could not participate in the session and was marked absent. However, for CBA sessions, the key instructions and exercises were presented by the Student Delivery System. If students arrived within about 5 minutes after other students started the exercises, the Test Administrators informed the student about the purpose of the test and allowed the student to begin.

For both CBA and PBA sessions, students were not allowed to leave the session unless it was absolutely necessary. If a student could not complete the session for any reason, the Test Administrator had to log the student out of the CBA session or collect the student's test material (PBA countries only). If the student was absent for more than 10 minutes from the test session (CBA or PBA), the Test Administrator recorded this student as "partially present" on the Session Attendance Form. Absences of 10 minutes or more in the questionnaire section did not affect the participant status of students.

For both CBA and PBA countries, Test Administrators were not allowed to provide any help with the test items. For CBA countries, the Test Administrator referred students who had questions to the "Help" function built into the Student Delivery System. However, they could answer questions about items in the Student Questionnaire following specific instructions in explanatory notes for Student Questionnaire items provided to them by the international contractors.

Observers were limited to necessary staff members and the PISA Quality Monitors. National Centres were responsible for ensuring that confidentiality arrangements were in place (see Standard 11.1 in Annex F). In most cases, it was national policy that an observer was required to sign a confidentiality agreement.

At the end of the computer-based administration (test, Student Questionnaire, and other international and national options), Test Administrators logged out any students still logged in to the test and collected and destroyed all logon forms. The Test Administrator then collected all USB sticks and conducted a quality-control check on the number of USB sticks and the information on the Student Tracking Form, Session Attendance Forms, and Session Report Form. Test Administrators also transmitted the test data following data-transmission procedures outlined by the National Centre. The assessment material from each administration session was then bundled together with the corresponding Session Attendance Forms and Session Report Form and shipped to the National Centre, typically within 24 hours of completing the assessment or follow-up session.

At the end of the paper-based administration (test, Student Questionnaire, and other international and national options), Test Administrators collected all assessment materials as well as the completed School Questionnaire from the School Co-ordinator. The assessment material from each administration session had to be bundled together with the corresponding Session Attendance Forms, Session Report Form, unused test booklets, and Student Questionnaires and shipped to the National Centre, typically within 24 hours of completing the assessment or follow-up session.

## **RECEIPT OF MATERIALS AT THE NATIONAL CENTRE AFTER TESTING**

It was recommended that the National Centre establish a database of schools before testing began to record the shipment of materials to and from schools, tallies of materials sent and returned, and to monitor the progress of the materials return, including completion of online questionnaires throughout the various steps in processing materials (for CBA countries).

It was recommended that upon receipt of materials back from schools, the counts of completed and unused booklets or USB sticks also be checked against the participation status information recorded on the Student Tracking Form.

## **MAIN SURVEY REVIEW**

NPMs were required to complete a structured review of their main survey operations. The review was an opportunity to provide feedback to the international contractors on the various aspects of the implementation of PISA and to provide suggestions for areas that could be improved. It also provided an opportunity for the NPM to formally document aspects such as the operational structure of the National Centre, the security measures that were implemented, the use of contractors for particular activities and so on.



The Main Survey Review Questionnaire was submitted online on a flow basis after the completion of each activity. The complete review questionnaire was due 4 weeks after the submission of the national database.

### **Notes**

1. For the remainder of this chapter, we will use the term “country” when referring to a country, economy, or adjudicated region.
2. Throughout this document, the terms “School Co-ordinator” and “Test Administrator” are used when discussing the administration of the test in schools. However, please note that some countries use School Associates, individuals who fulfil the role of both School Co-ordinator and Test Administrator. School Associates received a School Associate’s Manual and were trained by the National Centre.
3. In some countries the additional materials were supplied by schools.





7

## PISA quality monitoring

<b>Introduction .....</b>	112
<b>Field trial and main survey review questionnaires .....</b>	112
<b>National centre consultations .....</b>	112
<b>PISA quality monitor (PQM) visits .....</b>	113
<b>Data adjudication .....</b>	114

## INTRODUCTION

PISA data collection activities were undertaken in accordance with strict quality assurance procedures. The quality assurance procedure that ensures the PISA 2015 data are fit for use consists of two components: first, to develop and document procedures for data collection; and second, to monitor and record the implementation of those procedures. Chapter 6 describes the procedures which national centres were required to follow while this chapter considers the second part of the process – monitoring quality.

While the aim of quality control is to establish effective and efficient procedures and guide the implementation process, quality-monitoring activities were implemented to observe and record any deviations from those agreed procedures during the implementation of the survey. These activities included:

- Field Trial and Main Survey Review Questionnaires
- National Centre Quality Monitor (NCQM) visits and consultations
- PISA Quality Monitor (PQM) visits.

## FIELD TRIAL AND MAIN SURVEY REVIEW QUESTIONNAIRES

After the implementation of the field trial and the main survey, National Project Managers (NPMs) were asked to review and provide feedback to the international contractors on all aspects of their field operations. This information is used to guide future implementations of the assessment.

The Field Trial and Main Survey Review Questionnaires were organised around all aspects outlined in the NPM Manual:

- use of key documents and processes: use a rating system to review NPMs' level of satisfaction with the clarity of key documents and manuals
- communication with the international contractors
- review of the usefulness of the PISA Portal
- review of the quality of communication by activity
- implementation of national and international options: confirm if the National Centre had executed any national and international options as agreed
- review of the outcomes of and process for provision of national feedback on proposed test items
- security arrangements: review security arrangements to confirm they had been implemented
- sampling plan: confirm the PISA field trial and main survey tests were implemented as agreed in the sampling plan
- translation/adaptation/verification: review the translation, adaptation and verification processes to see if they were implemented in accordance with PISA technical standards and to a satisfactory level
- archiving of materials: confirm if the National Centre had archived the test materials in accordance with the technical standards
- printing: review the print quality agreement process
- test administration: review Test Administrators' training processes and test administration procedures
- quality assurance: review the PISA Quality Monitor (PQM) activity during the main survey implementation at the international level
- coding: review coder training procedures, coding procedures, coding designs and the time required for coding
- data management: review the data management processes, including student sampling, database adaptation, data entry, coding of occupational categories, validity reports, and data submission.

## NATIONAL CENTRE CONSULTATIONS

A large number of consultation meetings took place between senior staff of the international contractors and NPMs or other representatives of National Centres, in the context of NPM and training meetings. An extensive schedule of consultation meetings was developed prior to each meeting, and the consultations provided the opportunity for detailed discussion on a wide variety of PISA implementation matters on which additional advice or support was sought by the National Centre. In addition, the international contractors were in constant communication with all countries through email, Skype, webinars, and via the PISA Portal website.



## PISA QUALITY MONITOR (PQM) VISITS

The international contractor responsible for overseeing survey operations implemented all phases of the PISA Quality Monitor (PQM) process: interviewing (by phone and Skype) and hiring candidates in each of the countries, organising their training, selecting the schools to visit, and collecting information from the PQM visits.

PQMs are independent contractors located in participating countries who are hired by the international survey operations contractor. They visit a sample of schools to observe test administration and to record the implementation of the documented field operations procedures in the main survey. Typically, 2 to 3 PQMs were hired for each country, and they visited an average of 15 schools in each country. In countries with short test periods, up to 17 monitors were hired to ensure that on average 15 schools were observed in each country. If there were adjudicated regions in a country, it was usually necessary to hire additional monitors, as a minimum of 5 schools were observed in adjudicated regions.

All PISA Quality Monitors are nominated by the NPMs through a formal process of submission of nominations to the international survey operations contractor. Based upon the NPM nominations, which were accompanied by candidate resumes, the survey operations contractor selected monitors who were independent from the National Centre (not paid by or reporting directly to the NPM), knowledgeable in testing procedures or with a background in education and research, and able to communicate fluently in English. Where the resume did not match the selection criteria, further information or an alternate nomination was sought. In a few cases, a PQM did not meet one or more of the above criteria mainly because he or she was not fluent in English.

The PQM Manual, PQM self-training package, the national and international versions of the Test Administrator's Manual and script, and copies of data collection forms were made available to all monitors upon receipt of their signed confidentiality agreement via email and post. Self-training involved reading the materials and completing a quiz. The quiz was reviewed by survey operations staff who provided feedback on incorrect responses. After completing this self-study, PQMs were required to participate in two trainings: a webinar conducted by the survey operations contractor to review their role and responsibilities, and an in-country Test Administrator training conducted by the National Centre to familiarise monitors with national procedures and policies.

At the same time, the international survey operations contractor provided support and addressed any issues or concerns via email, telephone, or Skype. The PQMs and the international survey operations contractor collaborated to develop a schedule of test administration site visits to ensure that a range of different schools was covered and that the schedule of visits was both economically and practically feasible. The international survey operations contractor paid the expenses and fees directly to each monitor.

The School Co-ordinator<sup>1</sup> in each school was responsible for providing a link between the NPM and the school, its students, teachers, and principal, as well as organising a suitable venue for the testing. The international survey operations contractor supplied each PQM with a list of schools he or she was scheduled to monitor. This list included the contact information for the School Co-ordinator for each school so the PQM could obtain details for the test day.

The majority of school visits were unannounced to the Test Administrator. This, of course, was not possible where the Test Administrator and the School Co-ordinator were the same person (School Associate).

### Information collected in PQM visits during test administration

A Data Collection Form (DCF) was developed for PISA Quality Monitors to record their observations systematically during each school visit. The form covered the following areas:

- comparison of the adaptations to the English source versions of the school-level materials with the national language translations
- information about the National Centre's Test Administrator Training
- preparation for the assessment
- conducting the assessment
- general questions concerning the assessment.

PQMs recorded all key test session information using a hard copy of the Data Collection Form. After each session, the monitor entered the data from this form into the online version and submitted it to the international survey operations contractor. This form provided detailed data on test administration, including:

- session date and timing
- deviations from standard test procedures
- conduct of the students
- testing environment.

This information was used to check that the implementation in each school was in accordance with the PISA Technical Standards. The information was also called upon if a country's results showed, for example, a greater degree of country-item interaction.

## DATA ADJUDICATION

All quality assurance data collected throughout the cycle were entered and collated in a central data adjudication database. Comprehensive reports were then generated for the Technical Advisory Group (TAG) for consideration during the data adjudication process (see Chapter 14).

The TAG experts used the consolidated quality-monitoring reports from the central data adjudication database to make country-by-country evaluations on the quality of field operations, translation, school and student sampling, and coding. The final reports by TAG experts were then used for the purpose of data adjudication that took place in June 2016.

### Note

1. Throughout this document, the terms "School Co-ordinator" and "Test Administrator" are used when discussing the administration of the test in schools. However, please note that some countries use School Associates, individuals who fulfil the role of both School Co-ordinator and Test Administrator. School Associates received a School Associate's Manual and were trained by the National Centre.



8

## Survey weighting and the calculation of sampling variance

<b>Survey weighting .....</b>	116
<b>Calculating sampling variance .....</b>	123



Survey weights are required to analyse PISA data, to calculate appropriate estimates of sampling error and to make valid estimates and inferences of the population. The PISA Consortium calculated survey weights for all assessed, ineligible and excluded students, and provided variables in the data that permit users to make approximately unbiased estimates of standard errors, conduct significance tests and create confidence intervals appropriately, given the complex sample design for PISA in each individual participating country.

## SURVEY WEIGHTING

While the students included in the final PISA sample for a given country were chosen randomly, the selection probabilities of the students vary. Survey weights must be incorporated into the analysis to ensure that each sampled student appropriately represents the correct number of students in the full PISA population.

There are several reasons why the survey weights are not the same for all students in a given country:

- A school sample design may intentionally over or under-sample certain sectors of the school population: in the former case, so that they could be effectively analysed separately for national purposes, such as a relatively small but politically important province or region, or a sub-population using a particular language of instruction; and in the latter case, for reasons of cost, or other practical considerations, such as very small or geographically remote schools.<sup>1</sup>
- Available information about school size at the time of sampling may not have been completely accurate. If a school had a large student body, the selection probability was based on the assumption that only a sample of students from the school would participate in PISA. But if the school turned out to be small, all students would be included. In this scenario, there was a higher probability that the students would be selected in the sample than planned, making their inclusion probabilities higher than those of most other students in the sample. On the other hand, if a school, that was expected to be small, was actually large, the students included in the sample would have smaller selection probabilities than others.
- School non-response, where no replacement school participated, may have occurred, leading to the under-representation of students from that kind of school, unless weighting adjustments were made. It is also possible that only part of the PISA-eligible population in a school (such as those 15-year-old students in a particular grade) were represented by its student sample, which also requires weighting to compensate for the missing data from the omitted grades.
- Student non-response, within participating schools, occurred to varying extents. Sampled students who were PISA-eligible and not excluded, but did not participate in the assessment for reasons such as absences or refusals, would be under-represented in the data unless weighting adjustments were made.
- Trimming the survey weights to prevent undue influence of a relatively small subset of the school or student sample might have been necessary if a small group of students would otherwise have much larger weights than the remaining students in the country. Such large survey weights can lead to estimates with large sampling errors and inappropriate representations in the national estimates. Trimming survey weights introduces a small bias into estimates but greatly reduces standard errors (Kish, 1992).

The procedures used to derive the survey weights for PISA reflect the standards of best practice for analysing complex survey data, and the procedures used by the world's major statistical agencies. The same procedures were used in other international studies of educational achievement such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Studies (PIRLS), which were all implemented by the International Association for the Evaluation of Educational Achievement (IEA). The underlying statistical theory for the analysis of survey data can be found in Cochran (1977), Lohr (2010) and Särndal, Swensson and Wretman (1992).



Weights are applied to student-level data for analysis. The weight ( $W_{ij}$ ) for student  $j$  in school  $i$  consists of two base weights, the school base weight and the within-school base weight, and five adjustment factors, and can be expressed as:

### 8.1

$$W_{ij} = t_{2ij} f_{1i} f_{2ij} f_{1j}^A t_{1i} w_{2ij} w_{1i}$$

Where:

$w_{1i}$  (the school base weight) is given as the reciprocal of the probability of inclusion of school  $i$  into the sample;

$w_{2ij}$  (the within-school base weight) is given as the reciprocal of the probability of selection of student  $j$  from within the selected school  $i$ ;

$f_{1i}$  is an adjustment factor to compensate for non-participation by other schools that are somewhat similar in nature to school  $i$  (not already compensated for by the participation of replacement schools);

$f_{1j}^A$  is an adjustment factor to compensate for schools in some participating countries where only 15-year-old students who were enrolled in the modal grade for 15-year-old students were included in the assessment;

$f_{2ij}$  is an adjustment factor to compensate for non-participation by students within the same school non-response cell and explicit stratum, and, where permitted by the sample size, within the same high/low grade and gender categories;

$t_{1i}$  is a school base weight trimming factor, used to reduce unexpectedly large values of  $w_{1i}$ ; and

$t_{2ij}$  is a final student weight trimming factor, used to reduce the weights of students with exceptionally large values for the product of all the preceding weight components.

### The school base weight

The term  $w_{1i}$  is referred to as the school base weight. For the systematic sampling with probability proportional-to-size method used in sampling schools for PISA, this weight is the reciprocal of the selection probability for the school, and is given as:

### 8.2

$$w_{1i} = \begin{cases} I_g / MOS_i & \text{if } MOS_i < I_g \\ 1 & \text{otherwise} \end{cases}$$

The term  $MOS_i$  denotes the measure of size given to each school on the sampling frame.

The term  $I_g$  denotes the sampling interval used within the explicit sampling stratum  $g$  that contains school  $i$  and is calculated as the total of the  $MOS_i$  values for all schools in stratum  $g$ , divided by the school sample size for that stratum.

The measure of size ( $MOS_i$ ) was set as equal to the estimated number of 15-year-old students in the school ( $EST_i$ ), if it was greater than the predetermined target cluster size (TCS), which was 42 students for most countries that did a computer-based assessment, and 35 for most countries that did a computer-based assessment without adding collaborative problem solving or doing a paper-based assessment. For smaller schools the  $MOS_i$  value is given via the following formula, where again,  $EST_i$  denotes the estimated number of 15-year-old students in the school:

### 8.3

$$\begin{aligned} MOS_i &= EST_i && \text{if } EST_i \geq TCS; \\ &= TCS && \text{if } TCS > EST_i \geq TCS/2; \\ &= TCS/2 && \text{if } TCS/2 > EST_i > 2; \\ &= TCS/4 && \text{if } EST_i = 0, 1 \text{ or } 2. \end{aligned}$$

These different values of the measurement of size (MOS) are intended to minimise the impact of small schools on the variation of the weights, while recognising that the per student cost of assessment is greater in small schools.

Thus, if school  $i$  was estimated to have one hundred 15-year-old students at the time of sample selection then  $MOS_i = 100$ . And, if the country had a single explicit stratum ( $g = 1$ ) and the total of the  $MOS_i$  values of all schools was 150 000 students, with a school sample size of 150, then the sampling interval,  $I_1 = 150\ 000/150 = 1\ 000$ , for school  $i$  and others in the sample, giving a school base weight of  $w_{1i} = 1\ 000/100 = 10$ . Thus, the school should represent about 10 schools in the population. In this example, any school with 1 000 or more 15-year-old students would be included in the sample with certainty, with a base weight of  $w_{1i} = 1$  as the  $MOS_i$  is larger than the sampling interval. In the case where one or more schools have a  $MOS_i$  value that exceeds the relevant sampling interval value ( $I$ ), these schools become certainty selections, and the value of  $I$  is recalculated after removing them.

### The school base weight trimming factor

Once school base weights were established for each sampled school in the country, verifications were made separately within each explicit sampling stratum to determine if the school base weights required trimming. The school trimming factor ( $t_{1i}$ ) is the ratio of the trimmed to the untrimmed school base weight, and for most schools (and therefore most students in the sample) is equal to 1.0000.

The school-level trimming adjustment was applied to schools that turned out to be much larger than was assumed at the time of school sampling. Schools were flagged where the 15-year-old student enrolment exceeded  $3 \times \text{MAX}(TCS, MOS_i)$ . For example, if the target cluster size ( $TCS$ ) was 42 students, then a school flagged for trimming had more than 126 ( $= 3 \times 42$ ) PISA-eligible students, and more than 3 times as many students as was indicated on the school sampling frame. Because the student sample size was set at  $TCS$  regardless of the actual enrolment, the student sampling rate was much lower than anticipated during the school sampling. This meant that the weights for the sampled students in these schools would have been more than three times greater than anticipated when the school sample was selected. These schools had their school base weights trimmed by having  $MOS_i$  replaced by  $3 \times \text{MAX}(TCS, MOS_i)$  in the school base weight formula. This means that if the sampled students in the school would have received a weight more than three times larger than expected at the time of school sampling (because their overall selection probability was less than one-third of that expected), then the school base weight was trimmed so that such students received a weight that was exactly three times as large as the weight that was expected.

The choice of the value of three as the cut-off for this procedure was based on experience with balancing the need to avoid variance inflation, due to weight variation that was not related to oversampling goals, but to not introduce any substantial bias by altering many student weights to a large degree. Very few school weights were trimmed in any one country, and in most countries no school weights were trimmed.

### The within-school base weight

The term  $w_{2ij}$  is referred to as the within-school base weight. With the PISA procedure for sampling students,  $w_{2ij}$  did not vary across students ( $j$ ) within a particular school  $i$ . That is, all of the students within the same school had the same probability of selection for participation in PISA. This weight is given as:

#### 8.4

$$w_{2ij} = \frac{\text{enr}_i}{\text{sam}_i}$$

where  $\text{enr}_i$  is the actual enrolment of 15-year-old students in the school on the day of the assessment (and so, in general, is somewhat different from the  $MOS_i$ ), and  $\text{sam}_i$  is the sample size within school  $i$ . It follows that if all PISA-eligible students from the school were selected, then  $w_{2ij} = 1$  for all eligible students in the school. For all other cases  $w_{2ij} > 1$  as the selected student represents other students in the school besides themselves.

In the case of the grade sampling option, for direct-sampled grade students, the sampling interval for the extra grade students was the same as that for the PISA students. Therefore, countries with extra direct-sampled grade students (Iceland) have the same within school student weights for the extra grade students as those for PISA-eligible students from the same school.

Additional weight components were needed for the grade students in Germany and Italy. For these two countries, the extra weight component consisted of the class weight for the selected class(es) (all students were selected in the grade sample in the selected class(es)). In these two countries, the use of whole-classroom sampling for the grade samples resulted in the need for a separate weighting process.



## The school non-response adjustment

In order to adjust for the fact that those schools that declined to participate, and were not replaced by a replacement school, were not in general typical of the schools in the sample as a whole, school-level non-response adjustments were made. Within each country sampled schools were formed into groups of similar schools by the international sampling and weighting contractor. Then within each group the weights of the responding schools were adjusted to compensate for the missing schools and their students.

The compositions of the non-response groups varied from country to country, but were based on cross-classifying the explicit and implicit stratification variables used at the time of school sample selection. Usually, about 10 to 30 such groups were formed within a given country depending upon school distribution with respect to stratification variables. If a country provided no implicit stratification variables, schools were divided into three roughly equal groups, within each explicit stratum, based on their enrolment size. It was desirable to ensure that each group had at least six participating schools, as small groups could lead to unstable weight adjustments, which in turn would inflate the sampling variances. Adjustments greater than 2.0 were also flagged for review, as they could have caused increased variability in the weights and would have led to an increase in sampling variances. It was not necessary to collapse cells where all schools participated, as the school non-response adjustment factor was 1.0 regardless of whether cells were collapsed or not. However, such cells were sometimes collapsed to ensure that enough responding students would be available for the student non-response adjustments in a later weighting step. In either of these situations, cells were generally collapsed over the last implicit stratification variable(s) until the violations no longer existed. In participating countries with very high overall levels of school non-response after school replacement, the requirement for school non-response adjustment factors to all be below 2.0 was waived.

Within the school non-response adjustment group containing school  $i$ , the non-response adjustment factor was calculated as:

### 8.5

$$f_{1i} = \frac{\sum_{k \in \Omega(i)} w_{ik} \text{enr}(k)}{\sum_{k \in \Gamma(i)} w_{ik} \text{enr}(k)}$$

where the sum in the denominator is over  $\Gamma(i)$ , which are the schools,  $k$ , within the group (originals and replacements) that participated, while the sum in the numerator is over  $\Omega(i)$ , which are those same schools, plus the original sample schools that refused and were not replaced. The numerator estimates the population of 15-year-old students in the group, while the denominator gives the size of the population of 15-year-old students directly represented by participating schools. The school non-response adjustment factor ensures that participating schools are weighted to represent all students in the group. If a school did not participate because it had no PISA-eligible students enrolled, no adjustment was necessary since this was considered neither non-response nor under-coverage.

Table 8.1 shows the number of school non-response classes that were formed for each country, and the variables that were used to create the cells.

## The grade non-response adjustment

Because of perceived administrative inconvenience, individual schools may occasionally agree to participate in PISA but require that participation be restricted to 15-year-olds in the modal grade for 15-year-old students, rather than all 15-year-old students. Since the modal grade generally includes the majority of the population to be covered, such schools may be accepted as participants rather than have the school refuse to participate entirely. For the part of the 15-year-old population in the modal grade, these schools are respondents, while for the rest of the grades in the school with 15-year-old students, such a school is a refusal. To account for this, a special non-response adjustment can be calculated at the school level for students not in the modal grade (and is automatically 1.0 for all students in the modal grade). No countries had this type of non-response for PISA 2015, so the weight adjustment for grade non-response was automatically 1.0 for all students in both the modal and non-modal grades, and therefore did not affect the final weights.

If the weight adjustment for grade non-response was needed (as it was in earlier cycles of PISA in a few countries), it would have been calculated as follows:

Within the same non-response adjustment groups used for creating school non-response adjustment factors, the grade non-response adjustment factor for all students in school  $i$ ,  $f_{1i}^A$ , is given as:

### 8.6

$$f_{1i}^A = \begin{cases} \frac{\sum_{k \in C(i)} w_{1k} enra(k)}{\sum_{k \in B(i)} w_{1k} enra(k)} & \text{for students not in the modal grade} \\ 1 & \text{otherwise} \end{cases}$$

The variable  $enra(k)$  is the approximate number of 15-year-old students in school  $k$  but not in the modal grade. The set  $B(i)$  is all schools that participated for all eligible grades (from within the non-response adjustment group with school  $(i)$ ), while the set  $C(i)$  includes these schools and those that only participated for the modal responding grade.

This procedure gives, for each school, a single grade non-response adjustment factor that depends upon its non-response adjustment class. Each individual student has this factor applied to the weight if he/she did not belong to the modal grade, and 1.0 if belonging to the modal grade. In general, this factor is not the same for all students within the same school when a country has some grade non-response.

**Table 8.1 Non-response classes [Part 1/2]**

Country/ economy	Number of explicit strata*	Implicit stratification variables	Number of original cells	Number of final cells
Albania	13	ISCED2/Mixed/ISCED3 (3)	31	14
Algeria	12	ISCED level (4); Gender (3)	29	16
Argentina	6	Funding sector (2); Education level (4); Urbanicity (2); Secular/Religious (2)	30	15
Australia	49	Geographic Location (3); School gender composition (3); School socio-economic Level (11); ISCED level (3)	450	58
Austria	26	School type (4); Region (9); Programme (18); Percentage of girls (5)	241	29
Belgium	32	Grade repetition – Flemish and French Community (5), German Community and some Flemish and French Community (1); Percentage of Girls – Flemish and French Community (4), German Community and some Flemish and French Community (1); School type – French Community (4), German and Flemish Community (1)	199	35
Brazil	55	Public/Private (4); DHI Quintiles (6); ISCED level (5); Capital/Country (2); Urbanicity (2)	536	100
B-S-J-G (China)**	53	Selectivity (3); Funding (2)	90	31
Bulgaria	11	Type of school (8); Size of settlement (5)	123	27
Canada	98	Urbanicity (3); Funding (2); ISCED level (3)	191	51
Chile	25	National test score level (4); % Girls (6); Urbanicity (2); Geographic zone (4)	206	27
Colombia	23	Urbanicity (2); Funding (2); Weekend school or not (2); Gender (5); ISCED Programme orientation (4)	160	33
Costa Rica	6	Track (2); Urbanicity (2); Shift (2); Region (27); ISCED level (3)	100	26
Croatia	7	Gender (3); Urbanicity (3); Region (6)	69	21
Cyprus <sup>1</sup>	8	Language (2); ISCED level (3)	14	8
Czech Republic	32	Regions (15); School gender composition (3)	141	51
Denmark	6	School type (8); ISCED level (4); Urbanicity (6); Region (6); FO Group (4)	148	42
Dominican Republic	18	Funding (4); Location (3); Shift (6); School size (4); Programme (3)	95	21
Estonia	4	School type (3); Urbanicity (2); County (15); Funding (2)	78	20
Finland	10	Regional State Administrative agencies (7); School type (7)	44	14
France	4	School type (4); Funding (2)	18	9
FYROM	4	Urbanicity (2)	8	5
Georgia	23	Location (2); Language (11)	53	16
Germany	18	State for SEN and vocational schools (17); School type for Normal Schools (6)	67	27
Greece	3	Geographical Area-Public Private (16); School type (3)	87	25
Hong Kong (China)	5	Student Academic Intake (4)	13	6
Hungary	6	Region (7); Mathematics Performance (6)	131	42
Iceland	32	Population Density (2)	32	10
Indonesia	3	Funding (2); School strata (3); Region (8)	95	45
Ireland	9	Socio-Economic Status Quartile (4); Percent of Female Students Born in 1999 (4)	69	28



Table 8.1 Non-response classes [Part 2/2]

Country/ economy	Number of explicit strata*	Implicit stratification variables	Number of original cells	Number of final cells
Israel	12	ISCED level (4); Group Size (3); SES (4); District (3)	65	26
Italy	65	Region (10); Type of School (2)	112	39
Japan	4	Level of proportion of students taking University/College Entrance Exams (4)	16	14
Jordan	6	Urbanicity (2); Gender (3); Level (2); Shift (2)	55	28
Kazakhstan	49	Region (14); Urbanicity (2); ISCED level (3); ISCED programme orientation (2); Funding (2)	130	29
Korea	3	Urbanicity level (3); School gender composition (3)	24	12
Kosovo	15	ISCED level (4)	35	30
Latvia	5	School type (5)	18	11
Lebanon	13	Language (3); Gender (3)	53	25
Lithuania	25	Language (4); Location – Lithuanian-only (1), Other language (4); School type – Lithuanian-only (1), Other language (5); Non-private/Private (2)	53	21
Luxembourg	6	School gender composition (3)	9	6
Macao (China)	10	Gender (3); School orientation (2)	19	12
Malaysia	9	School category (6); School type (16); Urbanicity (2); Gender (3); ISCED level (2)	33	13
Malta	9	Gender (3)	17	8
Mexico	6	School programme (7); Funding (2); Urbanicity (2)	44	19
Moldova	27	Funding (2); ISCED program orientation (6)	39	14
Montenegro	11	Gender (3)	18	14
Netherlands	3	School tracks--PRO, VMBO schools (6), HAVO, VWO schools (3), Private schools (1)	10	6
New Zealand	4	School decile (4); School authority (2); School gender composition (3); Urbanicity (2)	35	20
Norway	3	None	9	4
Peru	8	Region (26); Gender (3); School type (6)	107	27
Poland	3	Basic/Vocational (2); Private/Public (2); Locality size (4); Gender composition (3)	35	6
Portugal	50	School management (2); School Location (3); Curriculum (3)	121	30
Puerto Rico (USA) <sup>2</sup>	2	Grade span (5); Region/District (7); Location-minority status (9); Gender (1); State level within census region (1)	32	10
Qatar	6	Gender (3); Language (2); Level (5); Funding (2); Programme orientation (3)	32	12
Romania	2	Language (3); Urbanicity (2); LIC Type (3)	13	8
Russian Federation	42	School location (9); School type (3)	164	56
Scotland	7	Gender (3); Area Type (6)	33	12
Singapore	4	Gender (3)	6	5
Slovak Republic	9	Region (8); School type (7); Language (3); Group (combination of exam, ESCS, Management and REP) (163)	207	33
Slovenia	7	Location (5); Gender (3)	141	39
Spain	41	None	100	39
Sweden	8	Geographic LAN – ISCED3 schools (21), other schools (1); Responsible authority – ISCED3 schools (3), other schools (1); Level of immigrants – ISCED2 and ISCED2/ISCED3 schools (4), other schools (1); Income quartiles – ISCED2 and ISCED2/ISCED3 schools (4), other schools (1)	95	32
Switzerland	25	School type (22); Canton (26)	149	29
Chinese Taipei	13	Region (6); School Gender (3)	65	32
Thailand	16	Region (9); Urbanicity (2); Gender (3)	129	81
Trinidad and Tobago	22	Gender (3); Urbanicity (2)	44	26
Tunisia	18	ISCED level (3); Funding (2); % Repeaters (4)	78	31
Turkey	36	School type (10); Gender (3); Location (2); Funding (2)	104	16
United Arab Emirates	43	School gender (3); Language (2); ISCED (3); Programme (2)	136	55
United Kingdom (ex. Scotland)	96	School performance – England and Wales (6), Northern Ireland (1); Local authority – England (152), Wales (22), Northern Ireland (30)	360	64
United States	9	Grade span (5); Region/District (7); Location-minority status (9); Gender (4); State level within census region (17)	183	22
Uruguay	11	Geographical area (4); Gender (3)	43	15
Viet Nam	15	Region (6); Province (63); School type (5); Study commitment (2)	139	25

\* For details of the explicit stratification, see Table 4.1, in Chapter 4.

\*\* B-S-J-G (China) refers to the four PISA-participating China provinces: Beijing, Shanghai, Jiangsu and Guangdong.

1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

2. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

## The within school non-response adjustment

Within each final school non-response adjustment cell, explicit stratum and high/low grade, gender, and school combination, the student non-response adjustment  $f_{2i}$  was calculated as:

**8.7**

$$f_{2i} = \frac{\sum_{k \in X(i)} f_{1i} w_{1i} w_{2ik}}{\sum_{k \in \Delta(i)} f_{1i} w_{1i} w_{2ik}}$$

where

$\Delta(i)$  is all assessed students in the final school non-response adjustment cell and explicit stratum-grade-gender-school combination; and,

$X(i)$  is all assessed students in the final school non-response adjustment cell and explicit stratum-grade-gender-school combination plus all others who should have been assessed (i.e. who were absent, but not excluded or ineligible).

The high and low grade categories in each country were defined to each contain a substantial proportion of the PISA population in each explicit stratum of larger schools.

The definition was then applied to all schools of the same explicit stratum characteristics regardless of school size. In most cases, this student non-response factor reduces to the ratio of the number of students who should have been assessed to the number who were assessed. In some cases of a small (i.e. fewer than 15 respondents) cell (i.e. final school non-response adjustment cell and explicit stratum-grade-gender-school category combinations) sizes, it was necessary to collapse cells together, and then apply the more complex formula shown above. Additionally, an adjustment factor greater than 2.0 was not allowed for the same reasons noted under school non-response adjustments. If this occurred, the cell with the large adjustment was collapsed with the closest cell within grade and gender combinations in the same school non-response cell and explicit stratum.

Some schools in some countries had extremely low student response levels. In these cases it was determined that the small sample of assessed students within the school was potentially too biased as a representation of the school to be included in the final PISA dataset. For any school where the student response rate was below 25%, the school was treated as a non-respondent, and its student data were removed. In schools with between 25 and 50% student response, the student non-response adjustment described above would have resulted in an adjustment factor of between 2.0 and 4.0, and so the grade-gender cells of these schools were collapsed with others to create student non-response adjustments.<sup>2</sup>

For countries with extra direct grade sampled students (Iceland), care was taken to ensure that student non-response cells were formed separately for PISA students and the extra non-PISA grade students. No procedural changes were needed for Germany and Italy since a separate weighting stream was needed for the grade students.

## Trimming the student weights

This final trimming check was used to detect individual student weights that were unusually large compared to those of other students within the same explicit stratum. The sample design was intended to give all students from within the same explicit stratum an equal probability of selection and therefore equal weight, in the absence of school and student non-response. As already noted, poor prior information about the number of eligible students in each school could lead to substantial violations of this equal weighting principle. Moreover, school, grade, and student non-response adjustments, and, occasionally, inappropriate student sampling could, in a few cases, accumulate to give a few students in the data relatively large weights, which adds considerably to the sampling variance. The weights of individual students were therefore reviewed, and where the weight was more than four times the median weight of students from the same explicit sampling stratum, it was trimmed to be equal to four times the median weight for that explicit stratum. The trimming of student weights was a rare occurrence, happening in only about 15% of the countries, with only a few cases within any country.

The student trimming factor ( $t_{2ij}$ ) is equal to the ratio of the final student weight to the student weight adjusted for student non-response, and therefore equal to 1.0 for the great majority of students. The final weight variable on the data file is the final student weight that incorporates any student-level trimming. As in all previous PISA cycles, minimal trimming was required at either the school or the student levels.



## National option students

Other than class-based grade sampling, three countries had national option students, each of which required a separate weighting stream. The weighting stream followed all the usual weighting steps. Mexico had one state which was sampled separately from the Mexico national sample (the state was also covered in the Mexico national sample). Spain had its 17 adjudicated regions in its extra weighting stream. (The Spain national sample was a subsample of the adjudicated regional sample, with the addition of schools from the one non-adjudicated region.) The United States had separate weighting streams for each of Puerto Rico, Massachusetts public schools, and North Carolina public schools (Massachusetts and North Carolina were also covered in the United States national sample).

Several other countries also had national option students but in these cases, weighting was done along with the PISA students (Australia, Denmark) if weights were required, or not, if not required (Luxembourg).

## International options

For both financial literacy and the teacher questionnaire, no weights were required nor calculated, given the way the samples were selected and the way these data were analysed. The unweighted financial literacy response rates were calculated, as were those for the teacher questionnaire, to be used as quality indicators, if needed.

## CALCULATING SAMPLING VARIANCE

A replication methodology was employed to estimate the sampling variances of PISA parameter estimates. This methodology accounted for the variance in estimates due to the sampling of schools and students. Additional variance due to the use of plausible values from the posterior distributions of scaled scores was captured separately as measurement error. Computationally the calculation of these two components could be carried out in a single program, such as *WesVar 5* (Westat, 2007). The SPSS and SAS macros were also developed. For further detail, see *PISA Data Analysis Manual, 2<sup>nd</sup> edition* (OECD, 2009).

## The balanced repeated replication variance estimator

The approach used for calculating sampling variances for PISA estimates is known as balanced repeated replication (BRR), or balanced half-samples; the particular variant known as Fay's method was used. This method is similar in nature to the jackknife method used in other international studies of educational achievement, such as TIMSS, and it is well documented in the survey sampling literature (see Rust, 1985; Rust and Rao, 1996; Shao, 1996; Wolter, 2007). The major advantage of the balanced repeated replication (BRR) method over the jackknife method is that the jackknife is not fully appropriate for use with non-differentiable functions of the survey data, most noticeably quantiles, for which it does not provide a statistically consistent estimator of variance. This means that, depending upon the sample design, the variance estimator can be unstable, and despite empirical evidence that it can behave well in a PISA-like design, theory is lacking. In contrast the BRR method does not have this theoretical flaw. The standard BRR procedure can become unstable when used to analyse sparse population subgroups, but Fay's method overcomes this difficulty, and is well justified in literature (Judkins, 1990).

The BRR method was implemented for a country where the student sample was selected from a sample of schools, rather than all schools, as follows:

- Schools were paired on the basis of the explicit and implicit stratification and frame ordering used in sampling. The pairs were originally sampled schools, except for participating replacement schools that took the place of an original school. For an odd number of schools within a stratum, a triple was formed consisting of the last three schools on the sorted list.
- Pairs were numbered sequentially, 1 to  $H$ , with pair number denoted by the subscript  $h$ . Other studies and the literature refer to such pairs as variance strata or zones, or pseudo-strata.
- Within each variance stratum, one school was randomly numbered as 1, the other as 2 (and the third as 3, in a triple), which defined the variance unit of the school. Subscript  $j$  refers to this numbering.
- These variance strata and variance units (1, 2, 3) assigned at school level were attached to the data for the sampled students within the corresponding school.
- Let the estimate of a given statistic from the full student sample be denoted as  $X^*$ . This was calculated using the full sample weights.

- A set of 80 replicate estimates,  $X_t^*$  (where  $t$  runs from 1 to 80), was created. Each of these replicate estimates was formed by multiplying the survey weights from one of the 2 schools in each stratum by 1.5, and the weights from the remaining schools by 0.5. The determination as to which schools received inflated weights, and which received deflated weights, was carried out in a systematic fashion, based on the entries in a Hadamard matrix of order 80. A Hadamard matrix contains entries that are +1 and -1 in value, and has the property that the matrix, multiplied by its transpose, gives the identity matrix of order 80, multiplied by a factor of 80. Details concerning Hadamard matrices are given in Wolter (2007). The choice to use 80 replicates was made at the outset of the PISA project, in 2000. This number was chosen because it is “fully efficient” if the sample size of schools is equal to the minimum number of 150 (in the sense that using a larger number would not improve the precision of variance estimation), and because having too large a number of replicates adds computational burden. In addition the number must be a multiple of 4.
- In cases where there were 3 units in a triple, either one of the schools (designated at random) received a factor of 1.7071 for a given replicate, with the other 2 schools receiving factors of 0.6464, or else the one school received a factor of 0.2929 and the other 2 schools received factors of 1.3536. The explanation of how these particular factors came to be used is explained in Appendix 12 of the PISA 2000 Technical Report (Adams and Wu, 2002).
- To use a Hadamard matrix of order 80 requires that there be no more than 80 variance strata within a country, or else that some combining of variance strata be carried out prior to assigning the replication factors via the Hadamard matrix. The combining of variance strata does not cause bias in variance estimation, provided that it is carried out in such a way that the assignment of variance units is independent from one stratum to another within strata that are combined. That is, the assignment of variance units must be completed before the combining of variance strata takes place, and this approach was used for PISA.
- The reliability of variance estimates for important population subgroups is enhanced if any combining of variance strata that is required is conducted by combining variance strata from different subgroups. Thus in PISA, variance strata that were combined were selected from different explicit sampling strata and also, to the extent possible, from different implicit sampling strata.
- In some countries, it was not the case that the entire sample was a two-stage design, of first sampling schools and then sampling students within schools. In some countries for part of the sample (and for the entire samples for Cyprus\*, Iceland, Luxembourg, Macao (China), Malta, Qatar, and Trinidad and Tobago), schools were included with certainty into the sampling, so that only a single stage of student sampling was carried out for this part of the sample. In these cases instead of pairing schools, pairs of individual students were formed from within the same school (and if the school had an odd number of sampled students, a triple of students was formed). The procedure of assigning variance units and replicate weight factors was then conducted at the student level, rather than at the school level.
- In contrast, in one country, the Russian Federation, there was a stage of sampling that preceded the selection of schools. Then the procedure for assigning variance strata, variance units and replicate factors was applied at this higher level of sampling. The schools and students then inherited the assignment from the higher-level unit in which they were located.
- Procedural changes were in general not needed in the formation of variance strata for countries with extra direct grade sampled students (Iceland) since the extra grade sample came from the same schools as the PISA students. However, since all schools in Iceland were certainty schools, students within the schools were paired so that PISA non-grade students were together, PISA grade students were together and non-PISA grade students were together. No procedural changes were required for the grade students for Germany and Italy, since a separate weighting stream was needed in these cases.
- The variance estimator is then:

### 8.8

$$V_{BRR}(X^*) = 0.05 \sum_{t=1}^{80} \left\{ (X_t^* - X^*)^2 \right\}$$

The properties of BRR method have been established by demonstrating that it is unbiased and consistent for simple linear estimators (i.e. means from straightforward sample designs), and that it has desirable asymptotic consistency for a wide variety of estimators under complex designs, and through empirical simulation studies.

\* See note 1 under Table 8.1.



## Reflecting weighting adjustments

This description does not detail one aspect of the implementation of the BRR method. Weights for a given replicate are obtained by applying the adjustment to the weight components that reflect selection probabilities (the school base weight in most cases), and then re-computing the non-response adjustment replicate by replicate.

Implementing this approach required that the PISA Consortium produce a set of replicate weights in addition to the full sample weight. Eighty such replicate weights were needed for each student in the data file. The school and student non-response adjustments had to be repeated for each set of replicate weights.

To estimate sampling errors correctly, the analyst must use the variance estimation formula above, by deriving estimates using the  $t$ -th set of replicate weights. Because of the weight adjustments (and the presence of occasional triples), this does not mean merely increasing the final full sample weights for half the schools by a factor of 1.5 and decreasing the weights from the remaining schools by a factor of 0.5. Many replicate weights will also be slightly disturbed, beyond these adjustments, as a result of repeating the non-response adjustments separately by replicate.

## Formation of variance strata

With the approach described above, all original sampled schools were sorted in stratum order (including refusals, excluded and ineligible schools) and paired. An alternative would have been to pair participating schools only. However, the approach used permits the variance estimator to reflect the impact of non-response adjustments on sampling variance, which the alternative does not. This is unlikely to be a large component of variance in any PISA country, but the procedure gives a more accurate estimate of sampling variance.

## Countries and economies where all students were selected for PISA

In Iceland, Luxembourg, Macao (China), Malta, and Qatar, all PISA-eligible students were selected for participation in PISA. It might be unexpected that the PISA data should reflect any sampling variance in these countries, but students have been assigned to variance strata and variance units, and the balanced repeated replication (BRR) method does provide a positive estimate of sampling variance for two reasons. First, in each country there was some student non-response. Not all PISA-eligible students were assessed, giving sampling variance. Second, the intent is to make inference about educational systems and not particular groups of individual students, so it is appropriate that a part of the sampling variance reflect random variation between student populations, even if they were to be subjected to identical educational experiences. This is consistent with the approach that is generally used whenever survey data are used to try to make direct or indirect inference about some underlying system.

## Notes

1. Note that this is not the same as excluding certain portions of the school population. This also happened in some cases, but this cannot be addressed adequately through the use of survey weights.
2. Chapter 11 describes these schools as being treated as non-respondents for the purpose of response rate calculation, even though their student data were used in the analyses.

## References

- Adams, R. J. and M. Wu (eds.) (2002), *PISA 2000 Technical Report*, OECD Publishing, Paris.
- Cochran, W. G. (1977), *Sampling Techniques, 3rd edition*, John Wiley and Sons, New York, NY.
- Judkins, D.R. (1990), Fay's Method for Variance Estimation, *Journal of Statistics*, Vol. 6, pp. 223-229.
- Kish, L. (1992), "Weighting for unequal Pi", *Journal of Official Statistics*, Vol. 8/2, pp. 183-200.
- Lohr (2010), *Sampling: Design and Analysis, Second Edition*, Brooks/Cole, Boston, MA.
- OECD (2009), *PISA Data Analysis Manual, 2nd edition*, OECD Publishing, Paris.
- Rao, J.N.K. and J. Shao (1996), On Balanced Half-Sample Variance Estimation in Stratified Random Sampling, *Journal of the American Statistical Association*, Vol. 91, pp. 343-348.

- Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys", *Journal of Official Statistics*, Vol. 1/4, pp. 381-397.
- Rust, K. and J.N.K. Rao (1996), "Replication methods for analyzing complex survey data", *Statistical Methods in Medical Research: Special Issue on the Analysis of Complex Surveys*, Vol. 5, pp. 283-310.
- Särndal, C., B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York, NY.
- Wolter, K. (2007), *Introduction to Variance Estimation, Second Edition*, Springer, New York, NY.



## 9

# Scaling PISA data

<b>Overview</b> .....	128
<b>Data yield and data quality</b> .....	128
<b>The IRT models for scaling</b> .....	141
<b>Latent regression model and population modelling</b> .....	145
<b>Analysis of data with plausible values</b> .....	147
<b>Application of IRT and population models to PISA</b> .....	149

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.



## OVERVIEW

The test design for PISA was based on a variant of matrix sampling (using different sets of items and different assessment modes) where each student was administered a subset of items from the total item pool. That is, different groups of students answered different yet overlapping sets of items. That 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 modelled using the underlying commonalities among the items. This regularity can be used to characterise students as well as items in terms of a common scale, even if not all students take identical sets of items. This makes it possible to describe distributions of performance in a population or subpopulation and to estimate the relationships between proficiency and background variables.

To increase the accuracy of the measurement, PISA uses plausible values – which are multiple imputations – drawn from a posteriori distribution by combining the IRT scaling of the test items with a latent regression model using information from the student context questionnaire in a population model.

In the following section, an overview of the data yield, data preparation, and data quality is given. Then the population model used for PISA (IRT analysis, latent regression model and computation of plausible values) is described formally, followed by demonstrating its application to the PISA data describing the national and international item calibration, as well as the computation of plausible values. The procedures utilised for the linking, with the aim to obtain equivalent scales, are further described.

## DATA YIELD AND DATA QUALITY

Before data were used for scaling and population modelling, 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 could be confirmed and data could be approved for scaling.

### Targeted sample size, routing and data yield

#### **Targeted sample size**

The main survey assessment design for PISA 2015 covered the domains of science, reading and mathematics, as well as financial literacy as an optional domain, as computer- and paper-based designs. The computer-based design also included the collaborative problem-solving (CPS) domain. The computer-based design for countries that opted out of the collaborative problem solving assessment is described in Chapter 2 of this technical report. These designs required participating countries to sample a minimum of 150 schools representing their national population of 15-year-old students. Countries taking the computer-based assessment (CBA) with collaborative problem solving needed to sample 42 students from each of 150 schools for a total sample of 6 300 students, while countries taking the computer-based assessment without collaborative problem solving or the paper-based assessment (PBA) needed to sample 35 students from each of 150 schools for a total sample of 5 250. It is important to understand that 88% to 92% of students received a form that consists of four 30-minute clusters, or sets of tasks, assembled from two domains, resulting in one hour of assessment time per domain with a total of two hours of testing time per student. An additional 8% to 12% of students received forms consisting of four 30-minute clusters covering three of the four core domains; science was included in each of these forms (see Chapter 2 for more details).

#### **Data yield**

Table 9.1 shows the sample sizes and assessment languages for all 72 participating countries. Note that a student was only considered a “respondent” and included in the analysis if the student responded to at least half of the test items. When less than half of the test items were answered, the student had to respond to at least one test item and have at least one non-missing response to a part of context questionnaire items ST012 or ST013 (ST012 has 8 questions that ask about how many TV's cars, etc. are in the household; ST013 asks how many books are in the house).



[Part 1/2]

Table 9.1 Test mode, sample size per country and language

Country/economy	Language	Test mode	Financial literacy	N of subsample	N of schools	N total
Albania	Albanian	PBA		5 215	230	5 215
Algeria	Arabian	PBA		5 519	161	5 519
Argentina	Spanish	PBA		6 349	234	6 349
Australia	English	CBA/CPS	X	14 530	758	14 530
Austria	German	CBA/CPS		7 007	269	7 007
Belgium	Dutch French German	CBA/CPS	X	5 675 3 594 382	288	9 651
Brazil	Portuguese	CBA/CPS	X	23 141	841	23 141
B-S-J-G (China)*	Chinese	CBA/CPS	X	9 841	268	9 841
Bulgaria	Bulgarian	CBA/CPS		5 928	180	5 928
Canada	English French	CBA/CPS	X	15 444 4 614	759	20 058
Chile	Spanish	CBA	X	7 053	227	7 053
Colombia	Spanish	CBA/CPS		11 795	372	11 795
Costa Rica	Spanish	CBA/CPS		6 866	205	6 866
Croatia	Croatian	CBA/CPS		5 809	160	5 809
Cyprus <sup>1</sup>	English Greek	CBA/CPS		775 4 796	126	5 571
Czech Republic	Czech	CBA/CPS		6 894	344	6 894
Denmark	Danish	CBA/CPS		7 161	333	7 161
Dominican Republic	Spanish	CBA		4 740	194	4 740
Estonia	Estonian Russian	CBA/CPS		4 338 1 249	206	5 587
Finland	Finnish Swedish	CBA/CPS		5 534 348	168	5 882
France	French	CBA/CPS		6 108	252	6 108
FYROM	Albanian Macedonian Turkish	PBA		1 338 3 895 91	106	5 324
Georgia	Azerbaijani Georgian Russian	PBA		205 4 954 157	262	5 316
Germany	German	CBA/CPS		6 504	256	6 504
Greece	Greek	CBA/CPS		5 532	211	5 532
Hong Kong (China)	Chinese English	CBA/CPS		5 238 121	138	5 359
Hungary	Hungarian	CBA/CPS		5 658	245	5 658
Iceland	Icelandic	CBA/CPS		3 371	124	3 371
Indonesia	Indonesian	PBA		6 513	236	6 513
Ireland	English Irish	CBA		5 638 103	167	5 741
Israel	Arabian Hebrew	CBA/CPS		1 683 4 915	173	6 598
Italy	German Italian Slovenian	CBA/CPS	X	1 581 9 914 88	474	11 583
Japan	Japanese	CBA/CPS		6 647	198	6 647
Jordan	Arabian	PBA		7 267	250	7 267
Kazakhstan	Kazakh Russian	PBA		4 808 3 033	232	7 841
Korea	Korean	CBA/CPS		5 581	168	5 581
Kosovo	Albanian	PBA		4 826	224	4 826
Latvia	Latvian Russian	CBA/CPS		3 584 1 285	250	4 869
Lebanon	English French	PBA		1 850 2 696	270	4 546
Lithuania	Lithuanian Polish Russian	CBA/CPS	X	5 153 624 748	311	6 525

[Part 2/2]  
Table 9.1 Test mode, sample size per country and language

Country/economy	Language	Test mode	Financial literacy	N of subsample	N of schools	N total
Luxembourg	English French German	CBA/CPS		215 1 440 3 644	44	5 299
Macao (China)	Chinese English Portuguese	CBA/CPS		3 651 779 46	45	4 476
Malaysia	English Malaysian	CBA/CPS		1 433 7 428	225	8 861
Malta	English	PBA		3 634	59	3 634
Mexico	Spanish	CBA/CPS		7 568	275	7 568
Moldova	Romanian Russian	PBA		4 258 1 067	229	5 325
Montenegro	Serbian	CBA/CPS		5 665	64	5 665
Netherlands	Dutch	CBA/CPS	X	5 385	187	5 385
New Zealand	English	CBA/CPS		4 520	183	4 520
Norway	Bokmål Nynorsk	CBA/CPS		5 007 449	229	5 456
Peru	Spanish	CBA/CPS	X	6 971	281	6 971
Poland	Polish	CBA	X	4 478	169	4 478
Portugal	Portuguese	CBA/CPS		7 325	246	7 325
Qatar	Arabian English	CBA		7 341 4 742	167	12 083
Romania	Hungarian Romanian	PBA		414 4 462	182	4 876
Russian Federation	Russian	CBA/CPS	X	6 036	210	6 036
Singapore	English	CBA/CPS		6 115	177	6 115
Slovak Republic	Hungarian Slovak	CBA/CPS	X	402 5 948	290	6 350
Slovenia	Slovenian	CBA/CPS		6 406	333	6 406
Spain	Basque Catalan Galician Spanish Valencian		X	141 1 202 161 5 092 140	201	6 736
Sweden	English Swedish	CBA/CPS		71 5 387	202	5 458
Switzerland	French German Italian		CBA	1 307 3 531 1 022	227	5 860
Thailand	Thai	CBA/CPS		8 249	273	8 249
Chinese Taipei	Chinese	CBA/CPS		7 708	214	7 708
Trinidad and Tobago	English	PBA		4 692	149	4 692
Tunisia	Arabian	CBA/CPS		5 375	165	5 375
Turkey	Turkish	CBA/CPS		5 895	187	5 895
United Arab Emirates	Arabian English	CBA/CPS		7 436 6 731	473	14 167
United Kingdom	English Welsh	CBA/CPS		13 818 339	288	14 157
United States	English	CBA/CPS	X	5 712	177	5 712
Uruguay	Spanish	CBA/CPS		6 062	220	6 062
Viet Nam	Vietnamese	PBA		5 826	188	5 826
All Countries	N/A	N/A	N/A	N/A	17 429	509 032

\* B-S-J-G (China) data represent the regions of Beijing, Shanghai, Jiangsu, and Guangdong.

1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

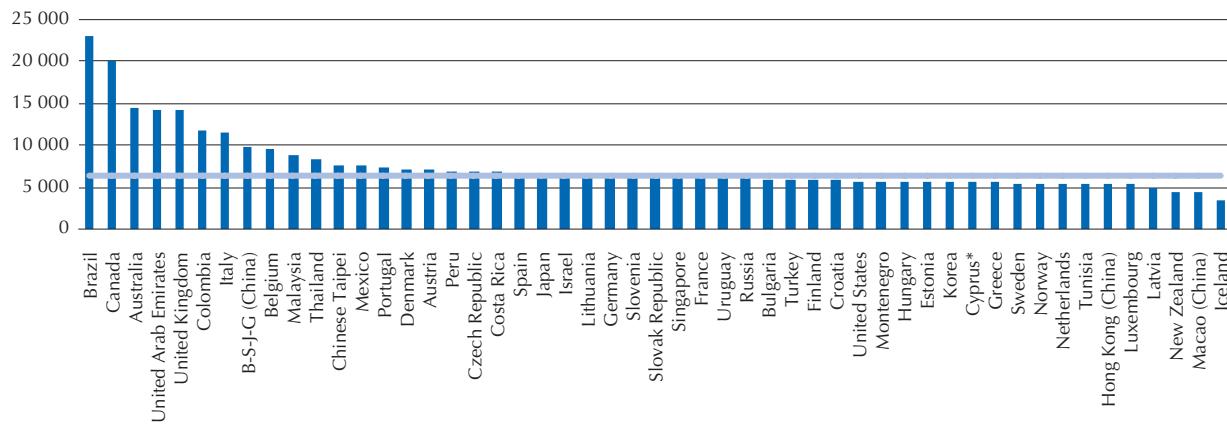
Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Note: Only students taking assessment in Dutch took financial literacy.

Due to population size and operational issues, not all countries satisfied the sample size requirement for the assessments they chose. Figures 9.1 and 9.2 show the sample yields for each participating country. Two charts are used because the sample size requirement is 6 300 for computer-based testing and collaborative problem solving and is 5 250 for both computer-based (without collaborative problem solving) and paper-based testing.

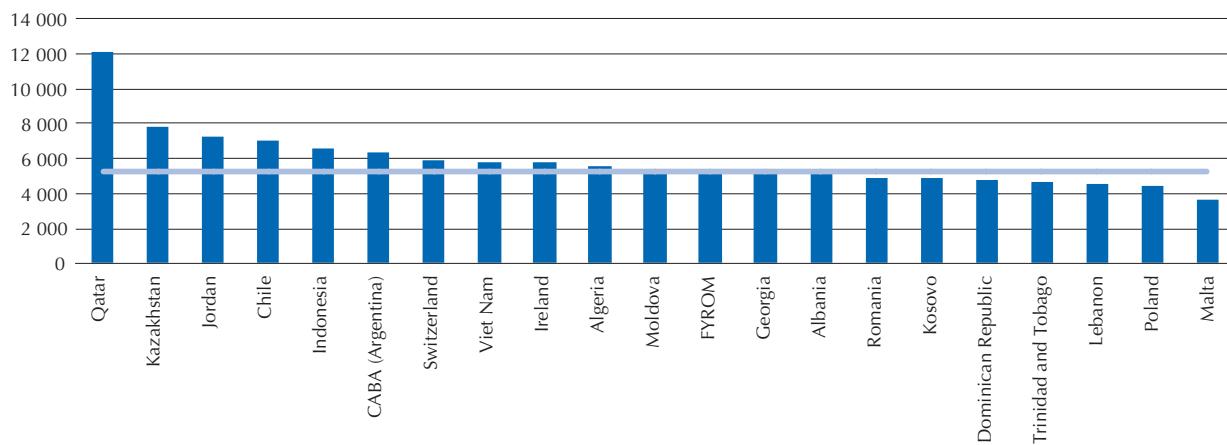


■ Figure 9.1 ■

**Sample yield for the participating countries with CBA/CPS format**

\* See note 1 below Table 9.1.

■ Figure 9.2 ■

**Sample yield for the participating countries with CBA or PBA format**

Since the sample sizes changed greatly from country to country, the numbers of schools and the sample sizes from each school changed as well. As seen in Table 9.1, number of schools runs from 44 (Luxembourg) to 841 (Brazil). But most countries met the requirement for the number of schools (a minimum of 150 schools).

**Classical test theory statistics: item analysis**

Item analyses were conducted on all computer and paper-based testing items at both the national and international levels to identify outliers, as well as human- and machine-scoring issues and other technical issues with regard to the CBA-collected data. All descriptive statistics were provided for observed responses as well as the various missing response codes and they were compared across modes and cluster positions for each item. 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 (that is the total score within a cluster) of students with specified response types for a given item

- point biserial correlations
- response time information within each domain per item and item cluster were examined in the PISA 2015 main survey.

Proportion correct and missing rates of trend items were compared to results from all prior PISA cycles when relevant. Statistics were compiled separately for the paper-based and computer-based assessments and also examined at the aggregate level across countries. The analyses were also performed separately for each country to identify outliers (single items that seem to work differently across assessment cycles and countries). Comparisons were made at a language-by-country level, and irregular cases, such as outliers as well as cases with obvious scoring rule deviations, were identified.

The PBA results included only paper-based student responses for the core domains of science, reading and mathematics (trend items only). The CBA results included computer-based student responses for the core domains of science, reading and mathematics (both trend and new items), as well as financial literacy and collaborative problem solving, where applicable. In addition, the results were disaggregated by language within a country (Note that *une-heure* (UH) booklet results are provided for countries where applicable).

**Table 9.2 Example output for examining response distributions**

**BLOCK M01 (UNWEIGHTED)**

**Response Analysis**

**A View Room**

ITEM 1	1	NOT RCH	OFF TSK	OMIT	0	1	TOTAL	R BIS =	0.5707
CM033Q01S	N	0	0	9	184	664	857	PT BIS =	0.4100
	Percent	0.00	0.00	1.05	21.47	77.48	100.00	P+ =	0.7748
TRN_MATH	Mean Score	0.00	0.00	0.89	2.55	5.47	4.79	DELTA =	9.98
	Std. Dev.	0.00	0.00	1.20	2.29	2.92	3.05		
	RESP WT	0.00	0.00	0.00	0.00	1.00		Item WT =	1.00

**Running Time**

ITEM 2	2	NOT RCH	OFF TSK	OMIT	0	1	TOTAL	R BIS =	0.6124
CM474Q01S	N	2	0	8	403	444	855	PT BIS =	0.4882
	Percent	0.23	0.00	0.94	47.13	51.93	100.00	P+ =	0.5193
TRN_MATH	Mean Score	0.00	0.00	2.25	3.28	6.24	4.80	DELTA =	12.81
	Std. Dev.	0.00	0.00	2.17	2.44	2.85	3.05		
	RESP WT	0.00	0.00	0.00	0.00	1.00		Item WT =	1.00

**Population Pyramids**

ITEM 3	3	NOT RCH	OFF TSK	OMIT	00	11	12	13	21	TOTAL	R BIS =	0.8725
DM155Q02C	N	10	0	227	163	71	59	11	316	847	PT BIS =	0.7445
	Percent	1.17	0.00	26.80	19.24	8.38	6.97	1.30	37.31	100.00	P+ =	0.4563
TRN_MATH	Mean Score	1.50	0.00	2.33	2.88	4.99	4.97	5.55	7.55	4.83	DELTA =	13.44
	Std. Dev.	1.12	0.00	1.63	1.99	1.98	2.13	2.46	2.26	3.05		
	RESP WT	0.00	0.00	0.00	0.00	0.50	0.50	0.50	1.00		Item WT =	2.00

Table 9.2 is an example of the response analysis output for a country using computer-based testing for the first three items in block/cluster M01. The first item, CM033Q01S, is the scored version of item CM033Q01 – a multiple-choice item. More details are given below for this item in the table.

The first column says CM033Q01S is the first item in the trend maths scale (TRN\_MATH).

In the second column, the first is the number of the item in the list, which is 1. All others are statistics for the response types, which are in the first row, starting from the third cell. They are:

1. N = Number of responses for the given type
2. Percent = Percent of responses for the given type
3. Mean Score = Mean score of the cluster (TRN\_MATH) for the given type
4. Std. Dev. = Standard deviation of the cluster (TRN\_MATH) 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 nor 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 = Wrong 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 2, Total is the sum of OMIT, 0 (Wrong) and 1 (Correct), i.e.  $855 = 8 + 403 + 444$ , which does not include NOT RCH, whose value is 2.

The statistics shown in the last two columns of Table 9.2 are ETS-developed indices. They are:

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. At ETS, it 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 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 sum of RESP WT values of all response type except NOT RCH.

Table 9.3 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 9.4). In this example, the third item is flagged for having an omit rate of greater than 10%, which prompts that further review is needed.

**Table 9.3 Example table providing summary item statistics**

<b>BLOCK M01 (UNWEIGHTED)</b>								
<b>Item Score Category Analysis (Partial credit model)</b>								
	Category	N	Pct. At	Pct. Below	Mean	Std. Dev.	Biserial	B *
ITEM 1	0	193	22.52	0.00	2.47	2.28		
CM033Q01S	1	664	77.48	22.52	5.47	2.92	0.5707	-1.3220
ITEM 2	0	411	48.07	0.00	3.26	2.44		
CM474Q01S	1	444	51.93	48.07	6.24	2.85	0.6124	-0.0788
ITEM 3	0	390	46.04	0.00	2.56	1.81		
DM155Q02C	1	141	16.65	46.04	5.02	2.09	0.6728	0.3992
	2	316	37.31	62.69	7.55	2.26	0.6133	-0.1780

<b>BLOCK M01 (UNWEIGHTED)</b>									
<b>Item Analysis Flag Summary</b>									
Item ID	Num Resp	Type	R BIS	P+	% NOTRCH	% OFFTSK	% OMIT	% MISS	Flags
CM033Q01	2	SCR	0.5707	0.7748	0.00	0.00	1.05	1.05	.....
CM474Q01	2	SCR	0.6124	0.5193	0.23	0.00	0.94	1.17	.....
DM155Q02	5	ECR	0.8725	0.4563	1.17	0.00	26.80	27.65	...O...

**Table 9.4 Flagging criteria for items in the item analyses**

Magnitude	Criteria for flagging items
Min rbis/rpoly	0.3
Min P+	0.2
Max P+	0.9
Omit % greater than	10
Off task % greater than	10
Not-Reached % greater than	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. Countries were therefore advised to use the latter statistics when evaluating the quality of items for their sample.

The PISA 2015 computer delivery platform successfully delivered, captured, and exported information for more than 900 items, with problems encountered in less than 1% of the items. Most of these items showed no obvious problems, yet there were a few items that had to be excluded from the analyses (in all countries/language groups) due to either almost no response variance, technical issues or very low item total correlations. These excluded items are shown in Table 9.5.

**Table 9.5 Items excluded from the IRT scaling based on classical item analyses or technical problems**

Domain	Item	Mode of administration
Maths (1 item)	CM192Q01	CBA
Science trend (7 item)	S327Q02/DS327Q02C*	PBA/CBA
	PS456Q01S	PBA
	PS456Q02S	PBA
	PS133Q01S	PBA
	PS133Q03S	PBA
	PS133Q04S	PBA
Collaborative problem solving (4 items)	CC104104	CBA
	CC104303	CBA
	CC102208	CBA
	CC105405	CBA

\*Five of the listed science items were dropped based on field-trial performance and content review. The items were not administered in the Microsoft computer-based instruments but were included in the paper-based assessments, as the booklets had been prepared before the decision was made to exclude the items. These items were excluded from the IRT scaling and population modelling. One item (DS327Q02C) was excluded from the main survey analysis, as it was discovered it had been dropped from the international analysis in 2003 and therefore could not be considered a trend item. Coders were instructed not to code this item and it was not included in the IRT scaling and population modelling. However, these six should have impacted the timing information on the clusters that contain them.

## Response time analyses

The computer-based platform captured response time information for all computer-based items. This information was used to compute the amount of time spent by the student on each item cluster at each cluster position within the spiral design. This information was also used to examine within- and between-country differences in response time and potential administration issues. The data for these analyses included item cluster response times and plausible values from the PISA 2015 main survey.

Detailed timing information is one of the two key features of the computer delivery platform (obviously) not available in paper-based assessments; another is process sequence information. Response times are recorded for each item in milliseconds; hence, they allow for precise, timing-related analyses. For instance, these data can be used to identify rapid guessing (e.g. Wise and DeMars, 2005) and/or potential administration issues (e.g. groups of students who take substantively longer to complete the assessment than expected). Timing information can also be used to address issues of speediness and fatigue, between-country differences in allocated time, position effects, and interaction effects with variables such as student performance. Sequence information, on the other hand, can provide insights into how students progress through a set of items, including the number of times that an aspect or an item component is revisited, item sets that are skipped, and items that are truly not reached. Further, sequence information can be used in conjunction with the timing data to identify potentially problematic items, units, and/or clusters.



Timing and process data were successfully recorded for all data collections in the CBA countries in the field trial and the main survey. The available timing data were instrumental in evaluating the level of student engagement and effort over the course of the four 30-minute clusters in addition to identifying response time outliers. Very little time spent on the items/assessment was interpreted as low effort; too much time spent on the items/assessment could be an indication of technical problems or low ability. Results from the analyses indicate that the CBA data provide valid information that can be used to evaluate student performance within and across countries.

### **Outliers**

Students were generally expected to complete each cluster within 30 minutes, but they had 60 minutes for the first two clusters and 60 minutes for the last two clusters with a break in between. In line with this expectation, an examination of the data shows that students rarely exceeded this maximum time. This was the case in the vast majority of cases; however, it was possible for some students to take additional time on the first and third clusters and less time on the second and fourth clusters, respectively, as the clusters were administered in pairs – before and after the mid-test break given at the 60-minute mark. Response times were identified greater than  $4.4478^*(\text{MAD})$  ( $\text{MAD} = \text{median}\{|x_i - \text{median}(x_j)|\}$ , where  $\{x_i\}$  is the collection of all sample values) above the sample mean within each cluster as outliers (Rousseeuw and Croux, 1993; Leys et al., 2013).

On average, 55 000 students took each cluster in the assessment; about 850 of them were labelled as outliers. Not surprisingly, all clusters have outliers. Table 9.6 shows the percentages of outliers by domain (science is split into science trend and science new).

**Table 9.6 Percentage of response time outliers in domains of PISA 2015 Main Survey**

Domain	Mathematics	Reading	Science trend	Science new	CPS*	FL**
Number of clusters	7	7	6	6	3	2
Percent of outliers	1.78%	1.89%	1.30%	1.21%	1.37%	2.49%

\* CPS = Collaborative problem solving

\*\* FL = Financial literacy

Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

### **Descriptive Statistics**

Table 9.7 presents descriptive statistics for the item cluster response times, by domain, with outliers excluded. These values are aggregated across countries and cluster positions. On average, students completed the items within each cluster in around 18 minutes, with 75% of the students completing the cluster in less than 22 minutes. With the outliers removed no student in any country took longer than 60 minutes to finish a given 30-minute cluster. Note that some variability in assessment time was expected as test administrators had to log off the computer-based assessment during the break one by one. Still, students who took close to one hour to complete a given 30-minute cluster would be unlikely to have had sufficient time to finish the subsequent cluster with which it was paired. That is, for the pair of clusters administered before or after the mid-test break, the use of up to 60 minutes for the first of the two clusters left no time to finish the second cluster. These long response times point to potential administration issues. On the other hand, there were also recorded cluster response times of less than one minute. It seems highly unlikely that a student could have completed a given cluster in under a minute; hence, this may indicate a technical problem with the data collection/time coding, or a breakoff, or input reflecting rapidly advancing through the items. It should be noted that 152 students had response times equal to 0 minutes due to technical issues (with 149 of these cases coming from Qatar); these values were excluded for all response time analyses.

**Table 9.7 Item cluster response time (in minutes) descriptive statistics**

Domain	Min	Q1	Median	Mean	Q3	Max	SD
Maths	0.95	13.53	17.38	17.40	21.25	36.93	5.88
Reading	0.81	13.47	17.09	17.18	20.86	36.79	5.78
Science trend	0.93	12.96	16.69	16.77	20.53	35.86	5.85
Science new	0.78	14.94	19.42	19.42	23.93	41.79	6.82
CPS*	3.04	19.24	22.52	22.77	26.18	42.14	5.62
FL**	1.17	14.77	19.28	19.12	23.82	38.70	6.56

\* CPS = Collaborative problem solving.

\*\* FL = Financial literacy.

Notes: Q1 is the 25th percentile and Q3 is the 75th percentile; all zero times were removed from the analyses. Argentina, Malaysia, and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

The median item cluster response time is similar across all domains for all countries taking the computer-based testing with the exception of collaborative problem solving, which is 3-5 minutes longer than the other domains. The standard deviation is almost the same across all domains, with science new and financial literacy items having slightly higher standard deviations.

To address the relationship between response time and student performance, median item response times grouped by proficiency levels were examined. Table 9.8 reports median response times by proficiency levels (both science and reading have Level 1a and 1b, instead of Level 1; both collaborative problem solving and financial literacy have only 5 levels). It is evident that the least able students (below Level 1) tended to complete a cluster in less time than other groups. Across all domains, more able students generally spent more time on each cluster. Except for collaborative problem solving, the differences between below Level 1 students and the highest level students exceeded around 7 minutes in all domains.

**Table 9.8 Cluster level response time by PV1 proficiency level (min)**

	Below Level 1	Level 1 <sup>1</sup>		Level 2	Level 3	Level 4	Level 5	Level 6
Mathematics	12.53		15.02	17.01	18.58	19.53	19.69	19.30
Reading	9.95	12.50*	15.22	17.20	18.12	18.32	18.20	17.96
Science trend	10.53	12.45	14.75	16.69	17.78	18.01	17.88	17.47
Science new	11.33	13.39	16.32	19.26	21.04	21.80	21.95	21.84
CPS	19.41		21.34	23.29	23.77	23.67	N/A	N/A
Financial literacy	14.88		19.38	21.33	22.52	23.17	N/A	N/A

1. Reading and science have 1a and 1b on Level 1.

Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

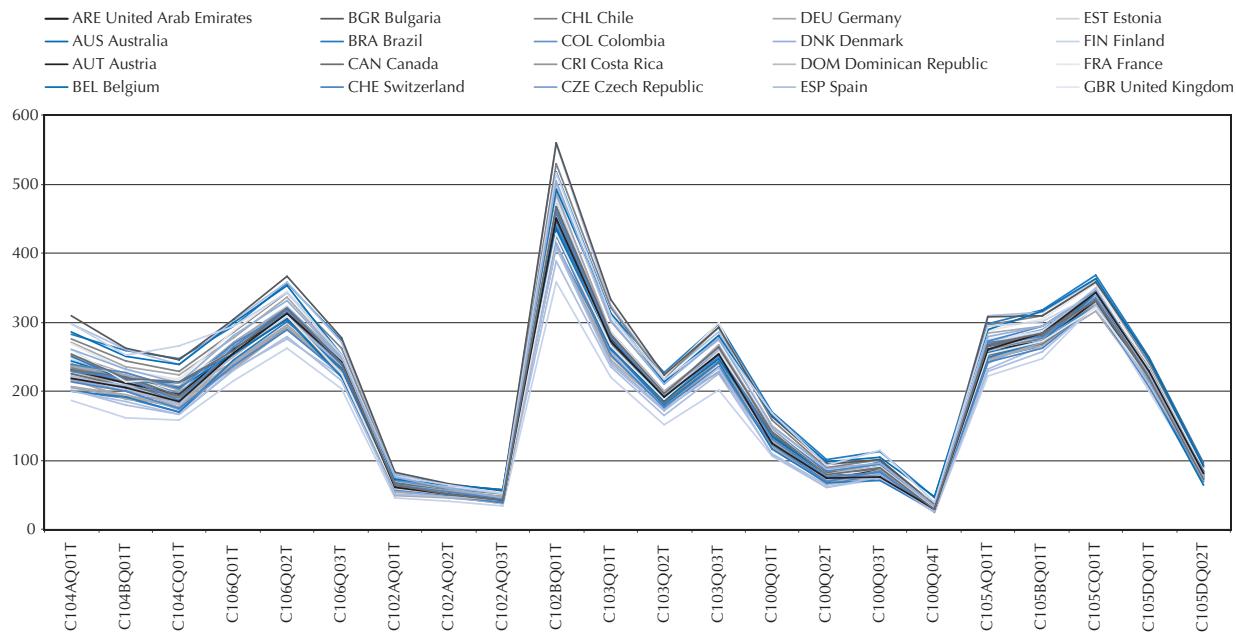
Response time was not only explored at the cluster level but also at the item level. The median response time for all items are similar across all countries. Figure 9.3 illustrates the median time of items across all countries using the CPS domain as an example.

Figures 9.4 and 9.5 show the median response time of science trend items and science new items based on the performance level across all countries (using weighted P+ and response times). The charts are sorted by the item response time. It can be seen that low performance students have almost identical response time patterns for both science trend items and science new items. The interaction between response time and ability (PV1) by items is greater for high performing students than for low performing students.

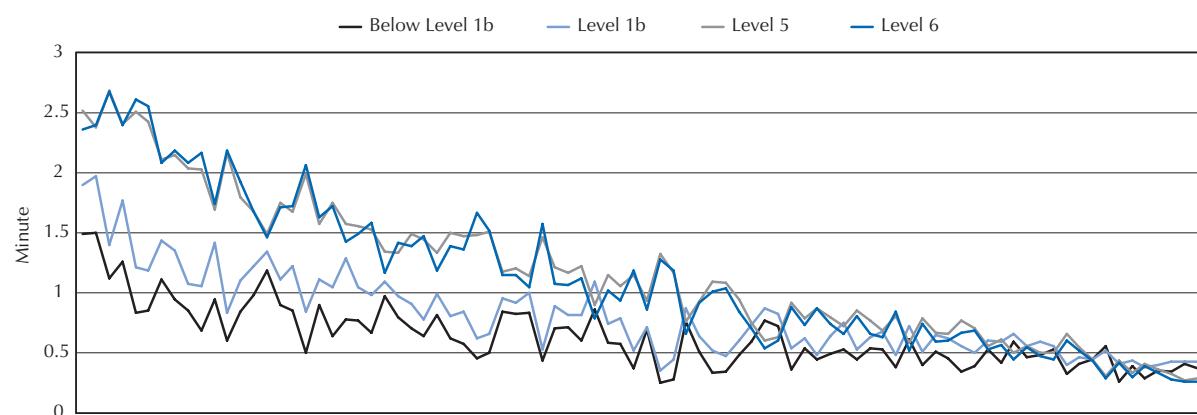
While the more able students generally need more time to complete the test, this is not true at the country level (see Figure 9.6). For example, Singapore has the highest average score in science, but its median response time is fairly close to the overall median time. Korea on the other hand has an unusually short median response time while its performance is relatively high.



■ Figure 9.3 ■

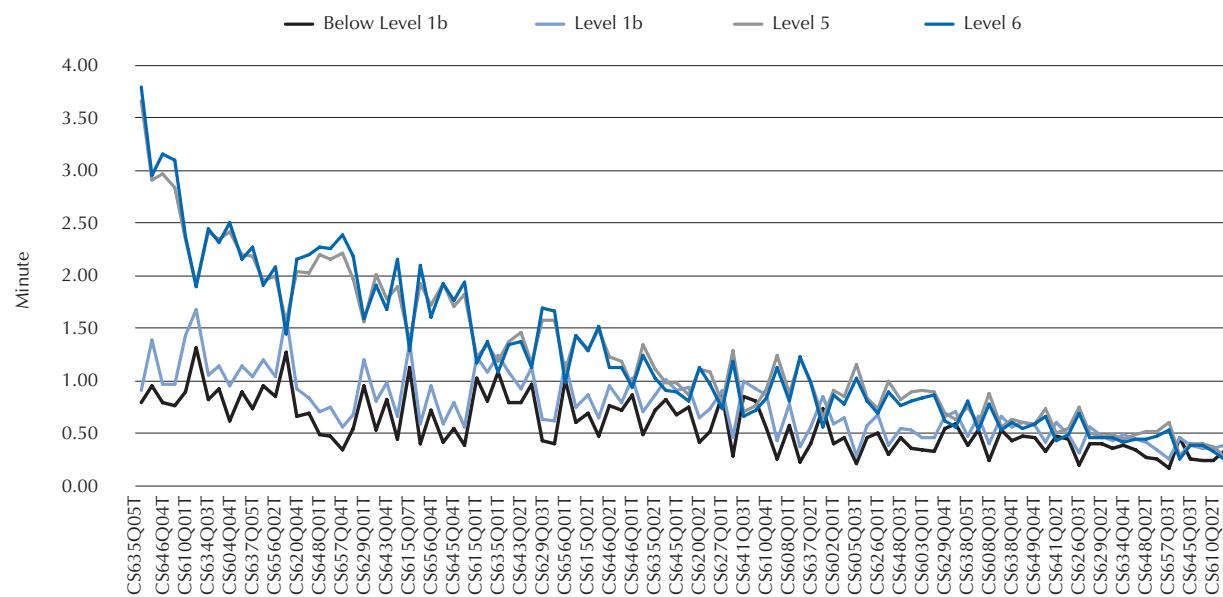
**Median response time by item – Collaborative problem solving**

■ Figure 9.4 ■

**Median response time by PV1 proficiency level – Science trend items**

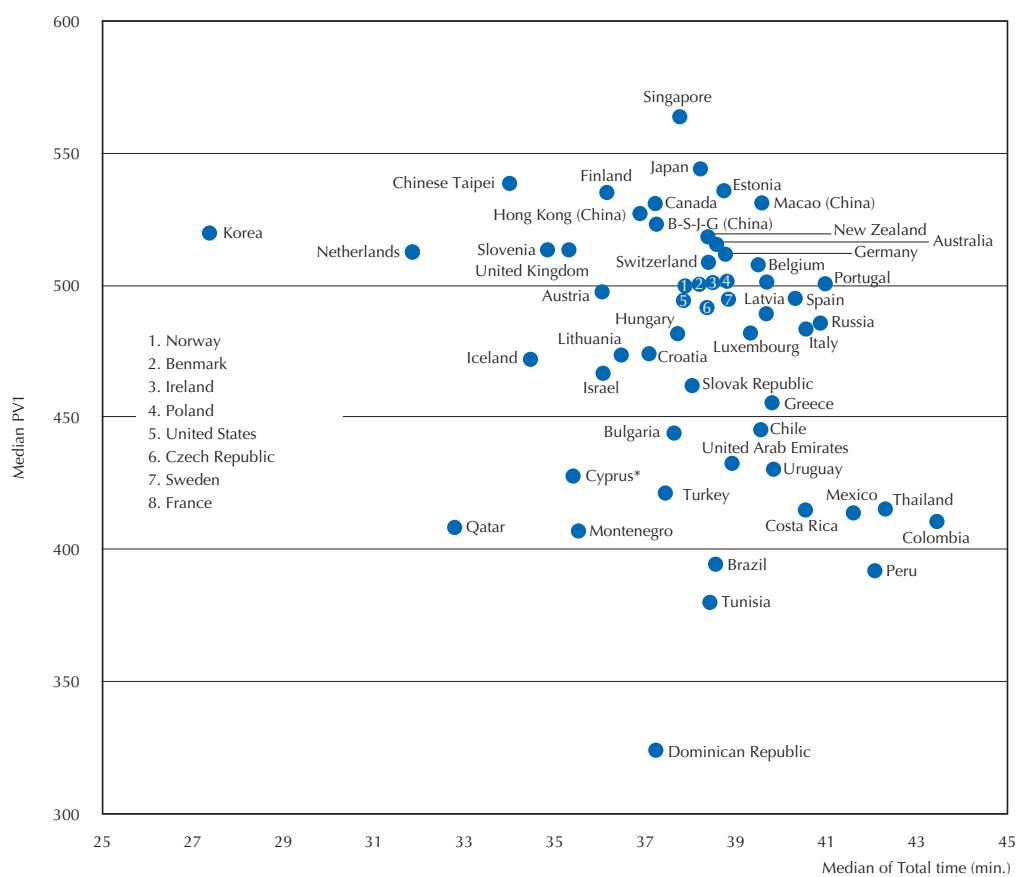
Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

■ Figure 9.5 ■

**Median response time by PV1 proficiency level – Science new items**

Note: Argentina, Malaysia, and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

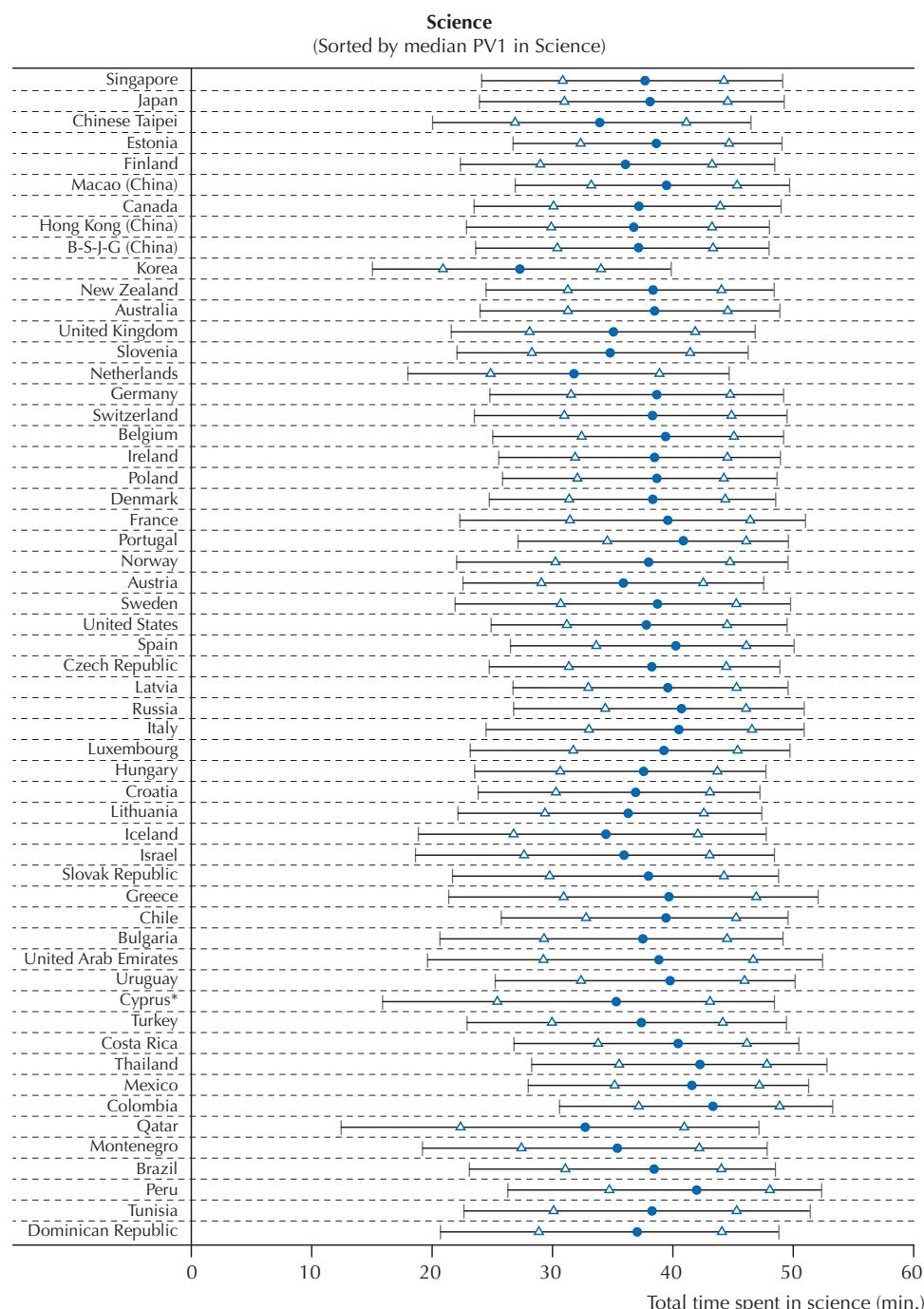
■ Figure 9.6 ■

**Median response time vs. country median score (PV1) – All science items (2 clusters)**



As part of this analysis, the within-country variability of response times was examined for all countries. Since science is the major domain for PISA 2015, with every student taking two clusters, results are presented for this domain only. Figure 9.7 shows the distribution of time spent on science for all countries sorted by their performance using the median of the first plausible value (PV1). The middle red solid dot is the median response time, and hollow triangles indicate the 25th and 75th percentiles of the response time, respectively, for a given country. The grey horizontal bars range from the 10th percentile of the response time to the 90th percentile of the response time for a given country. The figure suggests that the within-country variability is quite similar across countries.

■ Figure 9.7 ■  
**Variability of time used in science**



\* See note 1 below Table 9.1.

### **Administration (and possible student motivation) issues**

Results from the previous subsection suggest that there are few problematic patterns in the response times within and between countries. On average, students completed the entire test in 77.97 minutes ( $SD = 20.36$ ), with 1% of the students across countries taking longer than 120 minutes to complete the test. Some variability in assessment time was expected as test administrators had to log off the computer-based testing one by one. Students in Peru, Colombia, Thailand, and Tunisia took the longest median time to complete the test in 95.09, 90.12, 89.16, and 89.01 minutes, respectively. Students in Korea took the shortest median time to complete the test in 59.28 minutes.

There were five countries where 3% or more of the students exceeded the time limit: Tunisia (8.1%), Thailand (4.9%), United Arab Emirates (4.1%), Colombia (3.4%), and the Russian Federation (3.3%). On the other end of the distribution, 1.3% of the students completed the four clusters of the test in less than 30 minutes. These students were found in nearly all countries. The results for the students with very long or short total response times suggest that there were no systematic administration and/or motivation issues in specific schools. That is, in general, these students appear to be randomly distributed across schools and countries.

### **Position effects**

Item position effects are a common issue of concern in large-scale assessment programmes because substantial position effects can increase measurement error and introduce bias. The PISA 2015 main survey design balanced cluster position in order to control for the impact of item position and to monitor its impact of the item position on various item statistics. The cluster position effects were examined in terms of: 1) proportion of correct responses by cluster (average P+), 2) median response time by cluster and 3) rate of omitted responses by cluster (omission rate).

In order to establish a reference point for examining the magnitude of position effects, average P+ values were computed at the cluster level using both PISA 2009 and 2012 data. These values are shown in Table 9.9. We can see in this table that across the content domains there is a decrease of 0.04 to 0.08 points in the average P+ metric between cluster positions 1 and 4. For the PISA 2015 main survey data (see Table 9.10), the decrease is about 0.02 to 0.06 points in P+ values between cluster positions 1 and 4, which are smaller than the earlier cycles' values.

**Table 9.9 PISA 2009 and 2012 PBA proportion correct across clusters and across countries**

		Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1
2009	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	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

Note: Malaysia was not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

**Table 9.10 PISA 2015 CBA proportion correct across clusters and across countries**

	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1*
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
CPS	0.536	0.508	0.517	0.482	-0.054
FL	0.480	0.433	NA	NA	-0.047

\* For financial literacy, the difference is taken between positions 1 and 2 because these instruments only had two clusters.

Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

Table 9.11 shows the median cluster time averaged over all clusters at each position for all five domains. There are notable drops in median response times for all students from the first cluster to the second (3-6 minutes) and from the third cluster to the fourth (2-5 minutes); however, increases in the median response times for cluster 2 to cluster 3 (1-4 minutes) are relatively small compared to the drops. In addition to a decrease in P+ values from position 1 to position 4 for the 2015 main survey data (6-10%), there is a notable decrease in the median response times (around 4-6 minutes, i.e. nearly 20% reduction) for clusters administered in each of the four positions.



Table 9.11 PISA 2015 CBA median cluster timing averaged across countries (in minutes)

	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1*
Mathematics	19.81	16.91	17.34	15.71	-4.10
Reading	20.01	16.16	17.48	15.36	-4.65
Science trend	19.75	15.26	17.67	14.76	-4.98
Science new	23.38	17.40	20.73	16.89	-6.49
CPS	25.96	20.59	24.48	19.98	-5.98
FL	23.03	17.69	NA	NA	-5.33

\* For financial literacy, the difference is taken between positions 1 and 2 because these instruments only had two clusters.

Note: Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

The omission rates at different positions for all countries using computer-based assessments were analysed to further examine the quality of data affected by position. The omission rates for the PISA 2015 main survey in all domains and cluster positions are shown in Table 9.12. These rates do not include 'not reached' items.

Table 9.12 PISA 2015 CBA omission rates across clusters and across countries

	Position 1	Position 2	Position 3	Position 4	Position 4 - Position 1*
Mathematics	0.051	0.064	0.063	0.075	0.025
Reading	0.039	0.053	0.052	0.067	0.028
Science trend	0.029	0.046	0.038	0.052	0.023
Science new	0.027	0.039	0.035	0.045	0.018
FL	0.043	0.071	NA	NA	0.029

\* For financial literacy, the difference is taken between positions 1 and 2 because these instruments only had two clusters.

Note: Please note that Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

The omission rate for collaborative problem solving is 0% as students were forced to choose a response at each decision point in the tasks. Hence, omission rates for collaborative problem solving are not shown in the table.

Although no omission rate for any domain in any position exceeds 10%, the omission rates in Positions 2 and 4 are higher than those in Positions 1 and 3, respectively. Further, for reading, mathematics, and science, the omission rates in Position 3 are lower than those in Position 2, respectively. This is an indication that some students spent considerably more time on clusters 1 and 3, leaving them with less time for clusters 2 and 4.

## THE IRT MODELS FOR SCALING

### Moving from the Rasch model and partial credit model to the two-parameter logistic model and generalised partial credit model

The analysis of the PISA 2015 main survey data follows best practices outlined in, for example, Adams, Wilson, Glas and Verhelst (1995), Mislevy and Sheehan (1987), Yamamoto and Mazzeo (1992) and Wu (1997). More recent overviews of the different aspects of the methodology can be found in Glas and Jehangir (2014), Mazzeo and von Davier (2014), von Davier and Sinharay (2014), Weeks, von Davier and Yamamoto (2014), and von Davier (2006). 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., Adams, Skrondal and Rabe-Hesketh, 2004; von Davier and Yamamoto, 2007, 2004; Wu and Carstensen, 2007).

In prior PISA cycles (2000-2012), the Rasch model (1960) and the partial credit model (PCM; Masters, 1982) were used to estimate item difficulty parameters (calibrate/scale the items). The Rasch model is a mathematical model for the probability that an individual will respond correctly to a particular item, given the individual's location in a reference domain or dimension. The model postulates that the probability of response  $x$  to item  $i$  by a respondent depends on only two parameters, the difficulty of the item ( $\beta_i$ ) and the respondent's ability or trait level ( $\theta$ ), where:

**9.1**

$$P(x_i = 1 | \theta, \beta_i) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$



The probability of a positive response (e.g. solving an item) is strictly monotonically increasing in  $\theta$  and decreasing in  $\beta_i$ . If a respondent's ability matches the item difficulty, the expected probability of a correct response is equal to .50. Stated differently, item difficulty under the Rasch model can be interpreted as the location along the ability continuum at which a person is just as likely to answer the item correctly or incorrectly.

The partial credit model is an extension of the Rasch model to model the probability of responses to items with more than two ordered response categories. For a comprehensive review of the Rasch model, please refer to Chapter 3 (von Davier, 2016) of the *Handbook of Modern Item Response Theory* (2<sup>nd</sup> Ed.) edited by van der Linden (2016). For a review of the partial credit model, please refer to Chapter 7 of the same volume (Masters, 2016). Alternatively, von Davier and Sinharay (2014) review the use of IRT models in the context of international comparative assessments.

Concerns over the insufficiencies of the Rasch model to adequately address the complexity of the PISA data have been raised in the past (Kreiner and Christensen, 2014; Oliveri and von Davier, 2011, among others). Other national and international studies utilise more general IRT models (Mazzeo and von Davier, 2014; von Davier and Sinharay, 2014). The National Assessment of Educational Progress (NAEP), for example, uses the three-parameter IRT model and the generalised partial credit model (GPCM; Allen, Donoghue and Shoeps, 2001) as does the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) (Martin, Gregory and Stemler, 2000).

To address the concerns about usage of the Rasch model, PISA 2015 implemented the two-parameter-logistic model (2PLM; Birnbaum, 1968) for dichotomously scored responses and the generalised partial credit model (Muraki, 1992) for items with more than two ordered response categories.

The two-parameter logistic model 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  $\theta$  and the difficulty of the item difficulty,  $\beta_i$ . But in addition, the 2PLM allows that for every item, the association between this difference and the response probability can depend on an additional item discrimination parameter ( $\alpha_i$ ), characterising its sensitivity to proficiency. Under 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 follows:

## 9.2

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

where  $D$  is a constant of arbitrary size, often either 1.0 or 1.7, depending on the parameterisation used in the software implementation. Note that, for  $\alpha_i > 0.0$  this is a monotone increasing function with respect to  $\theta$ ; that is, the conditional probability of a correct response increases as the value of  $\theta$  increases. One important special case is when  $\alpha_i = 1.0/D$  for all items, in which case the Rasch model can be recognised as a special case of the two-parameter logistic model (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 the items.

A central assumption of the Rasch model, the two-parameter logistic model, and most IRT models is conditional independence (sometimes referred to as local independence). Under this assumption, item response probabilities depend only on  $\theta$  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. Moreover, the 2PLM assumes unidimensionality, that is, a single latent variable,  $\theta$ , that accounts for performance on the full set of items. This enables the formulation of the following joint probability of a particular response pattern  $x = (x_1, \dots, x_n)$  across a set of  $n$  items:

## 9.3

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



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:

**9.4**

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

The item parameter estimates obtained by maximising this function are used in the subsequent analyses.

The generalised partial credit model (Muraki, 1992), like the two-parameter logistic model, is a mathematical model for responses to items with two or more ordered response categories. While the two-parameter logistic model is suitable for dichotomous responses only, the generalised partial credit model can be used with polytomous and dichotomous responses. The generalised partial credit model reduces to the two-parameter logistic model when applied to dichotomous responses. For an item  $i$  with  $m_i + 1$  ordered categories, the model formula of the generalised partial credit model can be written as:

**9.5**

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

where  $d_i$  is the category threshold parameter.

The approach that was taken for the PISA 2015 analysis is a model that combines features of the Rasch model/partial credit model and the two-parameter logistic model/generalised partial credit model. This more general model was applied to the PISA 2015 field trial and main survey data. As a first step, the Rasch and partial credit models were applied to all trend items. The two-parameter logistic model or generalised partial credit model were used for items that showed poor fit to the Rasch model or partial credit model. Moreover, 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), as well as Oliveri and von Davier (2014, 2011) 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 to prior assessments that would otherwise be excluded from the trend measure (the more link items with good fit across groups, the more stable the link becomes).

In order to ensure that the IRT model used provides adequate fit to the observed data, different types of model checks are customarily applied. One of these checks is the evaluation of differential item functioning (DIF), which checks to determine whether items are harder or easier for a particular group compared to other groups of equal or similar ability. While the item parameters were estimated, empirical conditional percentage-correct statistics were monitored across the samples to test for differential item functioning between countries. More precisely, for each item, the empirical item characteristic curves (ICC) for each country-by-language group were compared to the expected ICC, given an estimate of the item parameter based on the total sample. If the empirical item characteristic curves for a certain group differed noticeably from the expected ICC, this would be evidence of differential item functioning. In order to examine the difference between the empirical and expected item characteristic curves, item fit statistics were calculated. More specifically, the approach for identifying differential item functioning in PISA 2015 is based on the mean deviation (MD) and the root mean square deviation (RMSD) fit statistics. Both measures quantify the magnitude and direction of deviations in the observed data from the estimated item characteristic curves for each single item. While mean deviation is more sensitive to deviations of observed item difficulty parameters from the estimated item characteristic curves, the root mean square deviation is sensitive to the deviations of both the item difficulty parameters and item slope parameters. In contrast to other measures for the evaluation of model data fit, such as INFIT and OUTFIT measures under the Rasch model, the mean deviation and root mean square deviation indices are not affected by sample size. Moreover, mean deviation and root mean square deviation statistics are available for a range of IRT models, while INFIT and OUTFIT measures are typically only provided for the Rasch model.

Group-specific item parameters (i.e. national item parameters) for items exhibiting group-level differential item functioning in the international calibration were estimated to reduce potential bias introduced by these deviations. This approach was favoured over dropping the group-specific item responses for these items from the analysis in order to retain the information from these responses. While the items with country differential item functioning 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-by-language group.

The software used for item calibration, *mdltm* (von Davier, 2005), implements an algorithm that monitored **differential item functioning** measures and that 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-by-language group or multiple country-by-language groups that showed the same level and direction of deviation.

### Measurement invariance (mode effect) model

Beginning in 2015, PISA became a computer-based assessment with a paper option for a small number of countries, while it was a paper-based assessment with optional computer-based scales in prior cycles. To address possible effects associated with this change, a mode effect study was conducted in the PISA 2015 field trial. The goal was to examine whether tasks presented in one mode (e.g. paper-based assessment) function differently when presented in another mode (e.g. computer-based assessment). A detailed description of the study and the results can be found in the section *Developing Common Scales for the Purpose of Trends* below. A comparison of different IRT models (extensions of the two-parameter logistic model assuming different mode effect parameters) in the field trial showed that the best fitting model is one that assumes item-specific mode effects for a subset of items, where items are affected differentially (i.e. some items could be more difficult, some could be at the same difficulty level, and some could become easier). This leads to a model that adds an item-specific effect for a subset of items to the difficulty parameter quantifying the item-specific difficulty difference between assessment modes, namely:

**9.15**

$$P(X = 1 | \theta, \alpha_i, \beta_i, \delta_m) = \frac{\exp(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_{mi})}{1 + \exp(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_{mi})}$$

Please note that this model is described again in the section *Developing Common Scales for the Purpose of Trends*; to avoid confusion the same numbering (9.15) is used in both sections. The computer-based difficulties are indexed with reference to the paper mode (computer-based items are indexed  $j = I + 1 \dots 2I$  and paper-based items  $i = 1 \dots I$ ). Then, difficulty parameters are decomposed into two components, that is,  $\beta_j = \beta_{i+1}$  with an optional mode effect parameter  $\delta_{mj}$  for  $j = i + 1$ , while it is assumed that the slope  $\alpha_j = \alpha_{i+1}$ . This decomposition is formulated so the difficulties are shifted by some item-dependent amount associated with the item or item feature. For other items, we may further assume that  $\delta_{mi} = 0$  (e.g. items for which the response mode differs but does not have a significant effect). As will be discussed below, for most items, there is no mode effect, that is  $\delta_{mj} = 0$ .

When the model given in formula (9.15) includes constraints across both modes on slope parameters, as well as potential constraints on the differential item functioning parameters  $\delta_{mi}$ , this establishes a measurement invariance (e.g. Meredith, 1993) IRT model that can be viewed as representing metric invariance. The more constraints of the type that  $\delta_{mj} = 0$  we have, the more we approach a model with strong or scalar invariance. Note that we already assume the equality of means and variances of the latent variable within groups in both modes because it is assumed that students receiving the test in computer or paper mode are randomly selected from a single population.

Using this model (9.15), it was possible to identify a subset of items that showed mode effects in the field trial. To account for these mode effects in the main survey, different item parameters were estimated for paired paper-based and computer-based items with substantive mode effects in the 2015 field trial; the paper-based and computer-based item parameters for items with no substantive mode effects were constrained to be the same (see *National and International Item Calibration and Handling of item-by-country/language and item-by-mode interactions* below for more information about the application of the IRT scaling approach to the PISA 2015 main survey data). This established an invariance model that assumes scalar or strong invariance for the majority of items and metric invariance for a minority of items for which difficulty differences were detected.



## LATENT REGRESSION MODEL AND POPULATION MODELLING

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 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 test 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 and that measure the same skill. Thus, individual achievement tests containing more than 70 items are common. Because the uncertainty associated with each estimated proficiency  $\theta$  is negligible, the distribution of proficiency or the joint distribution of proficiency with other variables can be approximated using individual proficiency estimates. 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 of any sort for the individual test taker. At the same time, ILSAs aim to achieve broad coverage of the tested constructs. The full set of items is organised into different, but linked, test forms; each individual receives only one booklet. 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 (small) number of responses to items in the booklet and responses to background questions administered in the PISA 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 estimation of proficiency distributions given responses to all domains received 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 PISA 2015 incorporated test responses (responses to the test 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 et al. 2006). For each student, 10 plausible values are computed. 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:

- 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 test items. The two-parameter logistic model is fitted for dichotomous item responses and the generalised partial credit model is fitted for polytomous item responses. Note that for a subset of trend items, the Rasch model and the partial credit model continue to be fitted for dichotomous and polytomous responses, respectively, to maintain consistency with prior PISA cycles.
- Population modelling using latent regressions and plausible value generation:* The population model assumes that item parameters are fixed at the values obtained in the calibration stage. Taking the item parameters estimates from Step 1, a latent regression model is fitted to the data to obtain regression weights ( $\Gamma$ ) and a residual variance-covariance matrix for the latent regression ( $\Sigma$ ). Next, 10 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  $\Gamma$  and  $\Sigma$  from the latent regression model.



3. *Variance estimation:* To obtain a variance estimate for the proficiency means of each country and other statistics of interest, a replication approach (see Johnson, 1989; Johnson and Rust, 1992; Rust, 2014) is used to estimate the sampling variability as well as the imputation variance associated with the plausible values.

As stated above, the population model used for PISA is a combination of the IRT model and a latent regression model. In the latent regression model, the distribution of the proficiency variable  $\theta$  is assumed to depend on the test 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 international large scale assessments. 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. For PISA it was decided to use the components for each country that accounted for 80% of the variance in order to avoid numerical instability due to potential overparameterization of the model. The use of principal components also serves to retain information for students with missing responses to one or more background variables. For the regression of the background variables on the proficiency variable it is assumed that:

## 9.6

$$\theta \sim N(\gamma\Gamma, \Sigma)$$

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

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

## 9.7

$$\theta_s = \gamma_{s0} + \gamma_{s1}y_1 + \dots + \gamma_{sL}y_L + \varepsilon_s$$

where  $\varepsilon_s$  is an error term for the assessment skill  $s$ .

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

## 9.8

$$\Sigma = \Theta\Theta^t - \Gamma(YY^t)\Gamma^t$$

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

## 9.9

$$P(\theta_j | x_j, y_j, \Gamma, \Sigma)$$

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

## 9.10

$$P(\theta_j | x_j, y_j, \Gamma, \Sigma) \propto P(x_j | \theta_j, y_j, \Gamma, \Sigma)P(\theta_j | y_j, \Gamma, \Sigma) = P(x_j | \theta_j)P(\theta_j | y_j, \Gamma, \Sigma)$$

where  $\theta_j$  is a vector of scale values (these values correspond to performance on each of the skills),  $P(x_j | \theta_j)$  is the product over the scales of the independent likelihoods induced by responses to items within each scale, and  $P(\theta_j | y_j, \Gamma, \Sigma)$  is the multivariate joint density of proficiencies of the scales, conditional on the principal components  $y_j$  derived from background responses, and parameters  $\Gamma$  and  $\Sigma$ . The item parameters are fixed and regarded as population values in the computation described in this section.



The basic method for estimating  $\Gamma$  and  $\Sigma$  using the expectation-maximization (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 formula above.

After the estimation of  $\Gamma$  and  $\Sigma$  is complete, plausible values are drawn from the joint distribution of the values of  $\Gamma$  for all sampled students in a three-step process. First, a value of  $\Gamma$  is drawn from a normal approximation to  $P(\Gamma, \Sigma | x_j, y_j)$  that fixes  $\Sigma$  at the value  $\hat{\Sigma}$  (Thomas, 1993). Second, conditional on the generated value of  $\Gamma$  (and the fixed value of  $\Sigma = \hat{\Sigma}$ ), the mean  $m_j^P$ , and variance  $\Sigma_j^P$  of the posterior distribution are computed using the same methods applied in the EM algorithm. In the third step, the  $\theta$  are drawn independently from a multivariate normal distribution with mean vector  $m_j^P$  and posterior co-variance matrix  $\Sigma_j^P$ . These three steps were repeated 10 times, producing 10 imputations of  $\theta$  for each sampled student.

The software DGROUP (Rogers et al., 2006) was used to estimate the latent regression model and generate plausible values (von Davier et al. 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.

## ANALYSIS OF DATA WITH PLAUSIBLE VALUES

If the scale proficiency values  $\theta$  were known for all students, it would be possible to directly compute any statistic  $t(\theta, y)$ , for example, a scale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient to estimate a corresponding population quantity  $T$ .

However, because the scaling models are latent variable models,  $\theta$  values are not observed. To overcome this problem, we follow the approach taken by Rubin (1987) and treat  $\theta$  as “missing” data. The value  $t(\theta, y)$  is approximated by its expectation given the observed data,  $(x, y)$ , as follows:

**9.11**

$$t^*(\bar{x}, \bar{y}) = E[t(\bar{\theta}, \bar{y}) | \bar{x}, \bar{y}] = \int t(\bar{\theta}, \bar{y}) p(\bar{\theta} | \bar{x}, \bar{y}) d\theta$$

It is possible to approximate  $t^*$  using plausible values (also referred to as imputations) instead of the unobserved  $\theta$  values. Plausible values are random draws from the conditional distribution of the scale proficiencies given the item responses  $x_j$ , background variables  $y_j$  and model parameters. For any student, the value of  $\theta$  used in the computation of  $t$  is replaced by a randomly selected value from the student’s conditional distribution. Rubin (1987) argues that this process should be repeated several times so that the uncertainty associated with imputation can be quantified. For example, the average of multiple estimates of  $t$ , each computed from a different set of plausible values, is a numerical approximation of  $t^*$  in the above formula; the variance among them reflects uncertainty due to not observing  $\theta$ . It should be noted that this variance does not include any variability due to sampling from the population.

It cannot be emphasised too strongly that the plausible values are not a substitute for test scores for individuals. Plausible values incorporate responses to test items and information about the background of responses; therefore, they cannot be used to compare individuals. Plausible values are only intermediary computations in the calculation of the integrals in the above formula in order to estimate population characteristics such as subgroup means and standard deviations. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated (von Davier, Gonzalez and Mislevy, 2009, provided examples and a more detailed explanation). The key idea lies in a contrast between plausible values and the more familiar ability estimates of educational measurement that are, in a sense, optimal for each student (e.g. bias corrected maximum likelihood estimates, which are consistent estimates of a student’s proficiency  $\theta$ , and Bayesian estimates, which provide minimum mean-squared errors with respect to a reference population). Point estimates that are optimal for individual students have distributions that can produce decidedly non-optimal (inconsistent) estimates of population characteristics (Little and Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects. For a further discussion of plausible values, see Mislevy et al. (1992).

After obtaining the 10 plausible values from the posterior distribution, they can be employed to evaluate formula (9.11) for an arbitrary function T as follows:

1. Use the first vector of plausible values (out of ten) for each student, calculate T as if the plausible values were the true values of  $\theta$ . Denote the result  $T_{1.}$
2. In the same manner as in step 1 above, estimate the sampling variance of T, or  $\text{Var}(T_{1.})$ , with respect to students' first vectors of plausible values. Denote the result  $\text{Var}_{1.}$
3. Carry out steps 1 and 2 for each of the  $U$  vectors of plausible values (in PISA 2015  $U=10$ ), thus obtaining  $T_u$  and  $\text{Var}_u$  for  $u = 2, \dots, U$ .
4. The best estimate of T obtainable from the plausible values is the average of the  $U$  values obtained from the different sets of plausible values:

### 9.12

$$T_{\cdot} = \frac{\sum_{u=1}^U T_u}{U}$$

5. An estimate of the variance of T is the sum of two components: an estimate of  $\text{Var}_u$  obtained as in step 4 and the variance among the  $T_u$ s:

### 9.13

$$\text{Var}(T_{\cdot}) = \frac{\sum_{u=1}^U \text{Var}_u}{U} + \left(1 + \frac{1}{U}\right) \frac{\sum_{u=1}^U (T_u - T_{\cdot})^2}{U - 1}$$

The first component in  $\text{Var}(T_{\cdot})$  reflects uncertainty due to sampling from the population; the second component reflects uncertainty due to measurement error, in other words because the students' proficiencies  $\theta$  are only indirectly observed through the item responses  $x$  and the background variables  $y$ .

#### **Example for partitioning the estimated error variance:**

The following example illustrates the use of plausible values in one particular country for partitioning the error variance. Tables 9.13 through 9.15 present data for six subgroups of students differing in the context questionnaire variable "books at home" (variable ST013Q01TA: 1 = 0-10 books; 2 = 11-25 books; 3 = 26-100 books; 4 = 101-200 books; 5 = 201-500 books; 6 = more than 500 books). Ten plausible values were calculated for each student in the science domain. Each column in this table presents the means of these 10 plausible values and the sampling standard error for each subgroup defined by the variable ST013Q01TA.

**Table 9.13 Example for use of plausible values to partitioning the error**

Plausible value	1		2		3		4		5		6	
	Mean	(s.e.)										
1	429.16	3.51	473.20	3.19	512.84	2.32	538.82	2.74	559.98	2.93	547.44	4.79
2	429.91	3.38	474.43	3.24	512.68	2.42	539.22	2.63	559.50	3.09	546.99	4.75
3	429.99	3.57	474.13	3.22	513.51	2.40	537.97	2.65	561.92	2.94	546.52	4.44
4	429.34	3.39	475.64	3.35	513.31	2.41	538.97	2.45	559.42	3.01	545.47	4.97
5	429.87	3.42	473.92	3.24	512.92	2.42	539.68	2.54	559.51	3.04	546.58	4.75
6	429.04	3.25	474.58	3.34	513.29	2.43	536.60	2.59	562.07	3.05	546.57	4.66
7	429.35	3.54	474.59	3.35	513.04	2.40	539.21	2.67	559.83	3.05	546.16	4.94
8	429.21	3.41	475.42	3.17	512.85	2.51	541.71	2.60	560.24	3.05	546.25	4.71
9	428.76	3.42	473.17	3.10	512.36	2.36	537.66	2.92	559.86	3.19	547.96	4.64
10	429.50	3.43	473.77	3.04	512.25	2.35	538.45	2.64	560.68	3.04	547.98	4.90

**Table 9.14 Example for use of plausible values to partitioning the error – sample error, measurement error and standard error based on the 10 PVs**

ST013Q01TA	Mean of 10 PVs	Sampling error	Measurement error	Standard error
1	429.41	3.43	0.43	3.46
2	474.29	3.23	0.87	3.34
3	512.90	2.40	0.42	2.44
4	538.83	2.64	1.42	3.00
5	560.30	3.04	1.02	3.20
6	512.90	2.40	0.42	2.44



The standard error reflects a component of error associated with the lack of precision of the measurement instrument and a component of error associated with sampling. The standard error can be reduced by either increasing the precision of the measurement instrument (for example, increasing the number of items) or reducing the sampling error. A resampling method is used to estimate the variance due to sampling. This component of variance is similar across the ten plausible values; the size is influenced by the homogeneity of proficiencies among students in the subgroup or by the precision of the survey instruments. The sampling error is smaller when the subgroup consists of students with similar proficiencies.

## APPLICATION OF IRT AND POPULATION MODELS TO PISA

This section describes the implementation of the different steps of IRT and population modelling using the PISA 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. More specifically, the procedures utilised for the linking, with the aim to obtain equivalent scales, are illustrated. It is also described how common scales were developed for the purpose of trends and an overview of the linking design and linking error is given.

Scaling and analyses of the PISA data were carried out separately for each of the domains: reading, mathematics, science, financial literacy and collaborative problem solving. 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 then carried out separately for each country.

### National and international item calibration

Item calibration is the first step in population modelling and provides the item parameters for the test items that are needed as one of the inputs for the population model used to calculate the plausible values. All analyses were carried out using the software *mdltm* (von Davier, 2005) for multidimensional discrete latent traits models. The software provides marginal maximum likelihood estimates obtained using customary expectation maximisation methods, with optional acceleration. Trend items were initially calibrated using the Rasch Model (Rasch, 1960) for dichotomous data and the partial credit model (Masters, 1982) for polytomous data by fixing the slope ( $a$ ) parameters to 1. Item fit was examined for all country-by-language-by cycle groups using a concurrent calibration. In cases of item misfit (root mean square deviation and mean deviation), the fixation of the slope parameters was released and the two-parameter logistic model (Birnbaum, 1968) for dichotomous data or the generalised partial credit model (Muraki, 1992) for polytomous data were estimated. In the case of new items the two-parameter logistic model and the generalised partial credit model were used for calibration. The result of the calibration is that all item parameters in each domain are located on a common scale.

Omitted responses prior to a valid response are treated as incorrect responses; whereas, omitted responses at the end of each of the two one-hour test sessions in both paper-based and computer-based assessments are treated as not reached/not administered. In the latter case, these responses have no impact on the IRT scaling. 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 sections *Population Modelling in PISA 2015* and *Generating Plausible Values*).

In total 83 maths items (83 items in the paper-based and 82 in the computer-based assessments), 103 reading items (in both paper- and computer-based assessments), 85 science items (in both paper- and computer-based assessments) and 43 financial literacy items (in the computer-based assessments only) were used as linking items between PISA 2015 and past PISA cycles. In addition, the PISA 2015 main survey contained 99 new science items and 121 collaborative problem solving items. Each domain was calibrated separately with a unidimensional IRT model. The item calibration included historical PISA data (PISA 2006–2012) in addition to the 2015 PISA data. This was done for the purpose of producing a linked scale for trend measurement reaching back to the last major domain cycle (in science 2006). Table 9.15 provides an overview of the distribution of the test items across the different PISA cycles and assessment modes (paper-based, computer-based) used for the calibration of PISA 2015.

Table 9.15 Distribution of the test items across PISA cycles and assessment modes by domain used in PISA 2015 item calibration (main survey)

		2006 only	2009 only	2012 only	2015 only	Items linked through 2 cycles	Items linked through 3 cycles	Items linked through 4 cycles	Total items in calibration across cycles	Total items in calibration across modes
Mathematics	PBA	12	–	26	–	52	2	30	122	82
	CBA	–	–	–	82	–	–	–	82	
Reading	PBA	–	30	–	–	36	64	3	133	103
	CBA	–	–	–	103	–	–	–	103	
Science trend	PBA	23	–	–	5	27	–	53	108	85
	CBA	–	–	–	85	–	–	–	85	
Science new	CBA	–	–	–	99	–	–	–	99	NA

Note: Each item is counted only once to avoid duplication.

Altogether, data from 536 177 students for reading, mathematics, and science; 140 074 students for financial literacy; and 418 808 students for collaborative problem solving were available for the PISA 2015 international IRT calibration together with PISA data coming from past PISA cycles (2006–2012)<sup>1</sup>. During the item calibration, sample weights standardised to represent each country equally were used.

As the samples for each PISA cycle came from somewhat different populations with different characteristics, the calibration procedure needed to take into account the possibility of any systematic interaction between the samples and the items that were used to produce estimates of the item parameters and sample distributions. For this reason, a multiple-group IRT model using country-by-language groups over different cycles and assessment modes 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 (concurrent calibration). The moments of these distributions were updated for every step in the iterations of the item parameter estimation.

The item calibration was completed in two consecutive steps. First, the data from all participating countries in 2015 and from the 2006–2012 cycles were analysed in an international calibration under the assumption that the common item parameters are the same across all countries and administration cycles. To account for mode effects for a subset of items identified in the PISA 2015 field trial mode effect study, different item parameters were estimated for the paired paper- and computer-based assessments; the item parameters for items in which no mode effects were found were constrained to be the same between the paper-based assessments and computer-based assessments.

In the subsequent step, unique item parameters were estimated to account for specific deviations for a subset of items. This involved a close monitoring of the IRT scaling for item-by-group interactions (group refers to country-by-language-by-cycle groups across modes) and allowing group-specific item parameters only in instances where deviations were identified. The following section describes this scaling step and the handling of item deviations from the model in more detail.

### Handling of item-by-country/language and item-by-mode 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. As such, some items in some countries or country-by-language groups may function somewhat differently from how the item generally functions in the majority of countries or groups. The same issue occurs when changing modes from a paper-to computer-based assessment or when comparing items across different assessment cycles over years. Some items may function differently in different assessment modes or in different cycles. For this reason, an analysis step was added that investigates item-by-country, item-by-cycle, and item-by-mode interactions, to identify cases in which an item may exhibit such deviant functioning in one or more groups.

The consistency of item parameter estimates across groups and countries was of particular interest to achieve common and unbiased measures of proficiencies that are comparable across countries, assessment modes, and assessments over time. 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, item-by-mode, or item-by-cycle interactions) as determined by group-specific item-fit statistics (mean deviation, MD; and root mean square deviation, RMSD), unique item parameters were estimated in a stepwise procedure. By allowing unique item parameters for items that show item-by-group interactions – in contrast to excluding such items, or forcing a common parameter – the measurement error is reduced without introducing bias. This approach



follows best practices described in the research literature on IRT and item fit assessment (Glas and Jehangir, 2014; Glas and Verhelst, 1995; Yamamoto, 1997; Oliveri and von Davier, 2014, 2011).

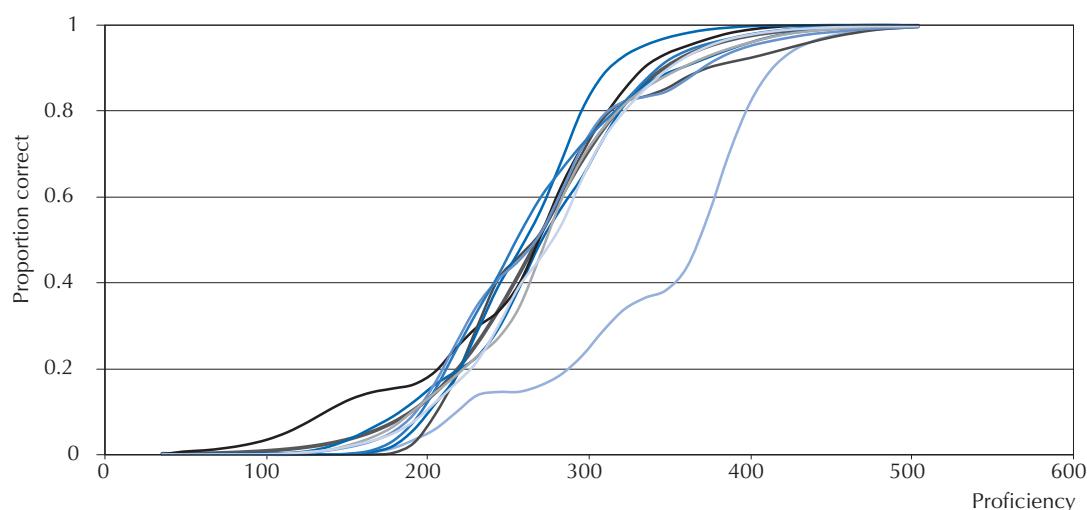
An algorithmic approach that automatically identified those group-by-item combinations requiring unique parameters based on differential item functioning detection was applied. Items not exhibiting appropriate fit using an international/common parameter received a group-specific parameter. However, if more than one group deviated from the international/common parameters in the same way (that is they showed similar differential item functioning), the algorithm assigned item parameters such that multiple groups share the same parameters, while differing from the international parameter estimate. For example, if two groups (e.g. two countries, or the same country in two PISA cycles) 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 (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 receive the same unique item parameter that is different from the international/common item parameter). If an item showed poor fit to a different extent in different groups, unique group-specific item parameters were used for further analysis. Thus, PISA allowed for different sets of item parameters to improve model fit and optimise the comparability of groups and countries.

To identify ill-fitting items, fit statistics were estimated using the mean deviation and the root mean square deviation (see *The IRT models for scaling* below for more information on these statistics). Poorly fitting items were revealed using a root mean square deviation  $> 0.12$  criterion and an mean deviation  $> 0.12$  and  $< -0.12$  criterion (a value of 0 indicates no discrepancy; in other words, a perfect fit of the model). The identification of poor fitting items and the replacement of international item parameters with group-specific (unique) parameters was carried out using an automatic algorithm in *mdlitm*. 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 one common scale.

In most cases, the item responses across groups and countries were accurately described by the international/common item parameters. For a subset of items, there was evidence of misfit for certain samples; however, this pattern was not consistent for any one particular group or country. Given this estimation and optimization approach, only a few items were dropped from the analysis in the PISA 2015 main survey. In all other cases, unique item parameters were estimated for items with substantial deviations from the international/common item parameters (poor fitting items). Figure 9.8 illustrates how the data from one group might not support the use of international item parameters.

■ Figure 9.8 ■

#### **Item response curve for an item where the international item parameter is not appropriate for one group (example from a different ILSA)**





The solid black line is the fitted two-parameter logistic item response curve that corresponds to the international item parameters; the other lines are observed proportions of correct responses at various points along the proficiency scale for the data from each subpopulation. The horizontal axis represents the proficiency scale. This plot indicates that the observed proportions of correct responses, given the proficiency, are quite similar for most countries and agree well with the IRT model-based curve. However, the data for one country indicated by the yellow line shows a noticeable departure from the common item characteristic curve. This item is far more difficult in that particular country, conditional on proficiency level. Thus, a unique set of item parameters was estimated for this country for this item.

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.

### **Mode effect study in the 2015 field trial: identifying items with mode effects**

To evaluate the stability of the link between paper- and computer-based assessments, a mode effect study was conducted with the PISA 2015f Field trial data where every country that later adopted a computer-based assessment in the main survey administered all trend items in both modes, thereby enabling a direct comparison between paper- and computer based assessment item parameters. The term “mode effect” refers to the observation that tasks presented in one mode (for example, paper-based) may function differently when presented in another mode (computer-based).

This section will first present a summary of the findings of the mode effect study and then illustrate in more detail the different approaches that were tested. In addition to some initial explorations (graphical model tests, correlations) of the similarity of item parameters across all domains, different formal conceptualisations of a “mode effect” were evaluated through statistical models (IRT model extensions) that contain parameters to quantify and compare potential differences between paper-based and computer-based assessments in an objective manner. This is followed by a description of how the best fitting model can be used to account and adjust for potential mode effects.

#### **Mode effect analyses and scaling approach for the main survey**

The mode effect study conducted in the PISA 2015 field trial showed that within mode, the item parameters are consistent across countries (and over time). Moreover, high correlations between item parameters across modes for all domains (0.94) was found. These findings indicate that the assessments administered in the two modes measure the same constructs. In the study with extended item response models that include different types of mode effect parameters, it was shown that the majority of items exhibit scalar or strong measurement invariance, while the remaining items exhibit metric invariance. Thus, a sound statistical link can be established, meaning computer-based and paper-based countries’ results can be reported on the same scales for 2015 and inferences about the scales are comparable.

For the subset of items with evidence of metric, but not scalar invariance, this meant that some items were somewhat harder while others were easier when delivered on the computer. That is, even among the subgroup that was identified and not fully invariant, the direction of the mode effect was not uniform. This finding discounted the hypothesis of a uniform mode effect that would somehow allow an overall scale adjustment.

For the subset of items that showed a difference of difficulty parameters between modes, separate item difficulties were calculated by mode. Slope parameters were the same across computer- and paper-based assessment modes.

Trend items that showed mode effects were identified in the field trial mode effect study. These items were re-examined in the main survey using population specific item-fit statistics (root means square deviation, mean deviation) in a concurrent calibration to confirm that the same invariance model can be applied to the main survey data. The items identified as exhibiting metric invariance were treated with mode-specific item difficulty parameters. Thus, possible mode effects are unlikely to impact the proficiency estimation, as the link between modes and cycles is established on a large number of trend items that show scalar (strong) invariance.

Chapter 12 provides information about which trend items are scalar invariant, sharing all characteristics across modes, and which items are partially or metric invariant, sharing a common slope parameter.



### **Graphical Model Tests and Correlations**

The comparison of mode differences in the current section is based on an approach that was first described by Rasch (1960). Parameter invariance across groups can be examined by applying the same identification constraints, and then estimating the parameters of a model in these groups separately and evaluating the level of agreement among the two sets of parameters. This “graphical model test” is useful to spot systematic differences between modes of administration, but it provides less statistical rigor than other model-based approaches. A graphical model test was conducted as a first step to examine the overall agreement of parameters of items administered in both modes and to explore potential drivers of any differences; the IRT models presented later (*IRT models to assess measurement invariance and mode differences*) were used to evaluate mode differences with a higher level of statistical rigor.

The PISA 2015 field trial incorporated an equivalent groups design that was implemented to aid the transition from paper- to computer-based assessment. This means that students were sampled in each country from a number of schools and then assigned randomly to one of two treatment conditions, taking the PISA field trial instruments on the computer or on paper. They were assigned independent of proficiency, prior experience, or other student variables.

This equivalent groups design allowed us to test the null hypothesis of “no mode effect”. The comparison was based on estimating parameters for the computer-based assessment mode and comparing them with parameters obtained from the (smaller) paper-based field trial sample, which was strengthened by combining it with data from prior paper-based PISA assessments ranging the 2000-2012 cycles. Due to the random assignment of students to modes, the underlying ability distributions of the paper- and computer-based field trial samples are assumed to be identical. As such, the computer-based parameters should not differ significantly, or systematically, from the parameters obtained in the 2000-2012 reanalysis (see *Developing common scales for the purpose of trends* later on in chapter) and verified using the paper-based field trial sample.

The following figures (9.9 and 9.10) show parameter comparisons between the mode-based samples. The IRT analyses for estimating these parameters are based on data from 68 field trial countries that submitted their data through November 2014 (reading, mathematics, and science: n = 150,983; financial literacy: n = 34,443).

Note that the paper-based item parameters were taken from the PISA 2000-2012 linking study that aimed at finding common parameters across five cycles of historical PISA data, and derived under the guiding principle of retaining as many Rasch model-based parameters as possible. More precisely, the paper-based item parameters were fixed to the estimates obtained from the linking study (where there were only paper-based assessment items), while the item parameters for the computer-based items were freely estimated (but constrained to be equal across countries). This was done simultaneously in the software *mdltm* (when fixing item parameters in a calibration, no additional constraints are needed since the fixation of parameters already takes care of the indeterminacy of the scale). Therefore, the paper-based set contains a number of slope values that are not estimated but fixed to 1 (retained Rasch Model items), which produces fewer pairs of freely estimated parameters. However, the difficulty parameters can be compared for all items that were administered in paper- and computer-based modes.

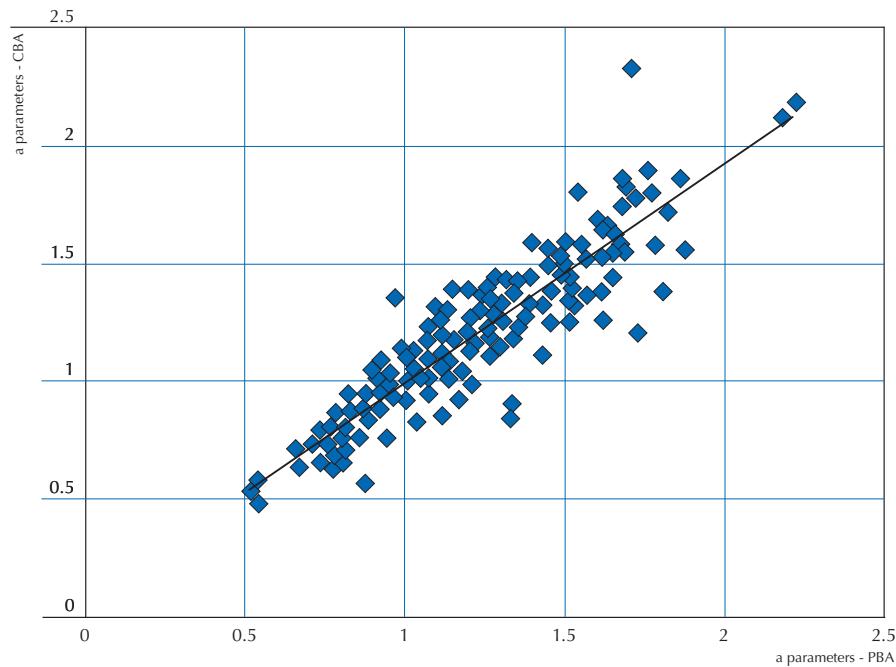
The distinction among the domains of reading, mathematics and science, as well as financial literacy was ignored because the parameters obtained across modes appeared to vary consistently across all domains.

These figures provide evidence of overall general agreement between the parameters based on the paper- and computer-based assessment modes. While there are differences, it appears that the level of difficulty of an item remains largely the same between paper-based parameters – based on historical data – and computer-based estimates. The same holds for the freely estimated slope parameters.

Moreover, correlations between the difficulty parameters for paper- and computer-based trend items are high within each domain, ranging from 0.92 to 0.95; the correlations between the discrimination parameters (slopes) range from 0.90 to 0.94 (note that only the two-parameter-logistic-model-based slopes were used to calculate correlations). The correlation of item difficulty parameters across modes and domains is 0.94, and the correlation of item slope parameters is 0.91. Table 9.16 presents an overview of these correlations. These high correlations as well as the Figures 9.9 and 9.10 suggest that the same constructs are being measured under both modes. The results from these field trial analyses suggested that a statistical link can be established whereby the computer- and paper-based countries’ results can be reported on the same scales for 2015 (for more information about the impact on mode effects on country means see *The impact of mode effects on country means in the field trial* later on in this chapter).

■ Figure 9.9 ■

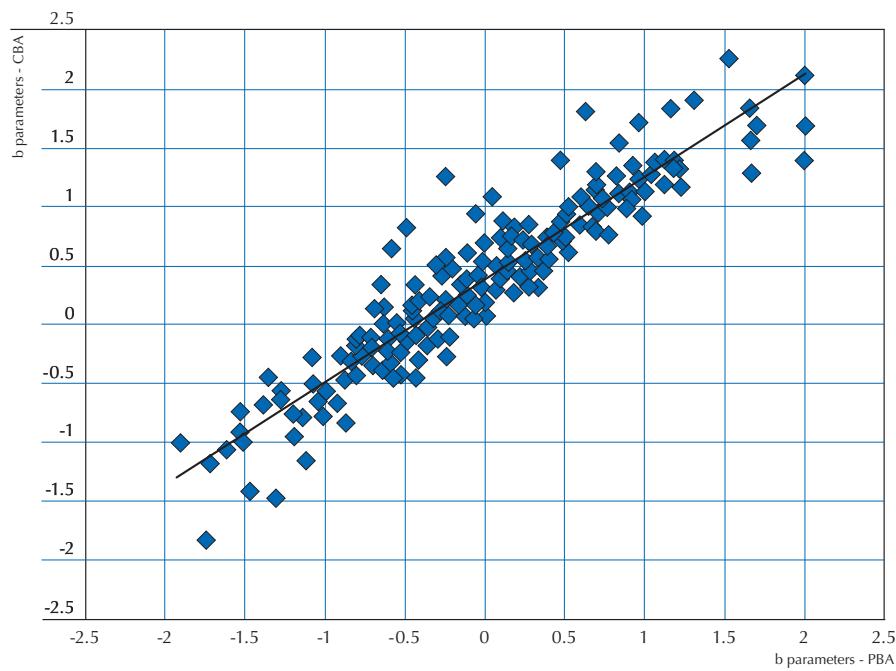
**Comparison of slope parameter estimates across paper-based (horizontal axis) and computer-based (vertical axis) assessment modes for the PISA 2015 field trial data**



Note: All domains with trend items (reading, mathematics and science, as well as financial literacy) are included.

■ Figure 9.10 ■

**Comparison of difficulty parameter estimates across paper-based (horizontal axis) and computer-based (vertical axis) assessment modes for the PISA 2015 field trial data**



Note: All domains with trend items (reading, mathematics and science, as well as financial literacy) are included.



**Table 9.16 Correlations of item difficulty and item slope parameters between paper-based and computer-based trend items within and across domains**

Domain	Correlation of difficulty parameters (PBA,CBA)	Correlation of slope parameters (PBA,CBA)
Mathematics	0.95	0.91
Reading	0.95	0.90
Science	0.92	0.94
Financial literacy	0.94	0.92
All Domains	0.94	0.91

### IRT models to assess measurement invariance and mode differences

Several mode-effect models that can be used to account for differences across groups were tested. More specifically, we tested whether mode differences are present on a global level, that is, whether the difference between paper and computer modes just adds or subtracts a level of difficulty to all assessment tasks, or whether the effect is person-specific, that is, whether some people are more affected by mode differences than others. Finally we tested a model that examines whether some items show mode effects, while others do not – that is, whether items are affected differently by mode effects.

Strong measurement invariance holds if the same item parameters fit the items independent of the mode of administration. A mode effect that homogeneously applies to all items in a test when changing the mode can be modelled by adding the same constant to all difficulty parameters in the case of the affected mode. Consider the two-parameter logistic model in formula (9.2) for greater ease of exposition. The notation in (9.2) can be transformed to the customary two-parameter logistic model notation via the transformation  $a = \alpha / 1.7$  and  $b = -\beta/\alpha$ .

If item  $i$  is presented in two different modes of administration, say paper and computer, a common (but arguably simple) assumption is that all items are “shifted” by a certain amount with respect to their difficulty. The reason for this could be that reading or, more generally, processing the item stem or stimulus is generally harder (by the same amount for all items and stimuli) on the computer, or entering a response on the computer is more tedious than filling in a bubble on an answer sheet of a paper-based instrument.

In order to represent this, we assumed a logistic IRT model with a general mode effect parameter –  $\delta_m$  that represents how much more difficult (or easy) solving an item is when presented in a given mode relative to a reference mode. For items presented in the reference mode, we assumed that model (9.2) holds; for items in the “new” mode, we assume that:

**9.14**

$$P(X = 1 | \theta, \alpha_i, \beta_i, \delta_m) = \frac{\exp(\alpha_i \theta + \beta_i - 1_{\{i>l\}} \delta_m)}{1 + \exp(\alpha_i \theta + \beta_i - 1_{\{i>l\}} \delta_m)}$$

The expression  $1_{\{i>l\}}$  denotes the indicator function which returns 1 if  $i > l$ . This shift by a mode effect in the same direction for all items in a specific mode can be thought of as a model with items (instead of items for each delivery mode separately) in which the difficulty parameters for items presented in one mode (say paper) are assumed to be  $\beta_i$  for  $i = 1, \dots, l$  and the item parameters for computer mode are appended as parameters  $j = l + 1, \dots, 2l$  and arranged in the same order and constrained to be  $\beta_j = \beta_{(j-l)} - \delta_m$ . That is, all computer-based item difficulties are simply shifted by a certain amount compared to paper-based items. Note that all IRT models illustrated in this section are based on the assumption of equivalent groups.

To explain why such an approach may be needed, or why it would be considered to estimate a mode effect in this way, the question of transitioning from paper- to computer-based testing can be used as a prototypical application. In such a setting, the same test items would exist in two modes, and information on how the test behaves (and more specifically, about the item parameters) may be available from large samples drawn from the reference population. In this setting, estimating completely new  $\beta_j$  may not be advisable, while estimating an overall mode effect –  $\delta_m$  could be considered for the purpose of adjusting for the effect of moving the items to computer administration.

In contrast to the assumptions made in model (9.14), one could argue that not all items become more difficult when moving them to the computer; some could be more difficult, some could be at the same difficulty level, and some could



even become easier. This leads to a model that adds an item-specific effect to the difficulty parameter. In model (9.15) we write this as a DIF parameter, which quantifies the difference from the paper-based assessment, namely:

### 9.15

$$P(X = 1 | \theta, \alpha_i, \beta_i, \delta_m) = \frac{\exp(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_{mi})}{1 + \exp(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\delta_{mi})}$$

As outlined above, the difference in comparison to the model of metric (or “weak”) factorial invariance (Meredith, 1993) is that the computer-based difficulties are written in reference to the paper mode and are decomposed into two components, that is,  $\beta_j = \beta_{i+I} - \delta_{mj}$ , while it is assumed that the slopes  $\alpha_j = \alpha_{i+I}$ . Again, this is written as a model with items, of which the first  $I$  items are presented in the reference mode, while the second  $I$  items are presented in the “new” mode. This decomposition is formulated so the difficulties are shifted by some item-dependent amount associated with the item or item feature. For paper-based items  $i \leq I$  we can assume  $\delta_{mi} = 0$ . In addition, there may be other items for which we may further assume that  $\delta_{mi} = 0$  (e.g., items for which the response mode differs but does not have a significant effect on item difficulty). These unaffected items are the basis for linking across modes, and below we show that these can indeed be assumed to be the majority of items.

The model given in formula (9.15) with constraints across both modes on slope parameters, as well as potential constraints on the DIF parameters, establishes a measurement invariance (e.g., Meredith, 1993) IRT model that can be viewed as representing a mixture of items with strong and weak factorial invariance. The more constraints of the type  $\delta_{mi} = 0$  we have, the more we approach a model with strong factorial invariance. Note that the equivalent groups design allows us to assume the equality of means and variances of the latent variable in both modes because it is assumed that students receiving the test via computer or paper mode are randomly selected from a single population.

Finally, if it cannot be assumed that the mode effect is a constant shift in difficulty for all students, one may assume that an additional ability  $\vartheta$  is required to predict the response probabilities in the new mode accurately. We still assume the same average in the paper- and computer-based ability distribution for the domain specific dimension; the additional mode dimension is independent. This leads to Model (9.16) in which a second latent variable was assumed, that is, another random effect was added to the item function for items administered in the new mode. The expression  $\alpha_{mi}\vartheta$  in Model (9.16) below indicates that there is a second slope parameter  $\alpha_{mi}$  for items administered in the new mode ( $i = I, \dots, 2I$ ) and that the effect of the mode is person dependent and quantified by the second latent variable  $\vartheta$ . We obtain:

### 9.16

$$P(X = 1 | \theta, \alpha_i, \beta_i, \delta_m) = \frac{\exp(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\alpha_{mi}\vartheta)}{1 + \exp(\alpha_i\theta + \beta_i - 1_{\{i>I\}}\alpha_{mi}\vartheta)}$$

Note that the slope parameters and item difficulties,  $\alpha_i$ ,  $\beta_i$ , are as before in models (9.14) and (9.15) equal across modes. Only the additional “mode slope” parameter  $\alpha_{mi}$  needs to be estimated for all items administered in the “new” mode, plus the joint distribution  $f(\theta, \vartheta)$  for which we can assume that the variables are uncorrelated, that is,  $\text{cov}(\theta, \vartheta) = 0$ .

In formula (9.16) it is assumed that the effect of the person “mode” variable varies across items, which is likely the more plausible variant, even though a mode in which person-dependent but item-homogenous effects  $\alpha_{mi}\vartheta$  (a Rasch variant of a random mode effect) could also be defined. Models (9.14), (9.15), and (9.16) can be applied to multiple populations, that is, by assuming one population per participating country or language group in PISA.

We conducted an empirical comparison of the models based on the field trial data. Table 9.17 below shows the results of models (9.14), (9.15), and (9.16) for a multiple population mode effects analysis using the PISA field trial data. All analyses were conducted with the software *mdltm* (von Davier, 2005). As a general rule, lower values for the statistics (Akaike information criterion, AIC; Bayesian information criterion, BIC; Consistent Akaike Information Criterion, CAIC, log-penalty, and Akaike) indicate better fit. However, when the magnitude of the statistics is similar, the more parsimonious model should be preferred. In all cases, Model (9.16) has the lowest values for these statistics, yet they do not differ appreciably from the fit for Model (9.15). To provide additional evidence for this interpretation we examined



the marginal reliability of scores under each model as well as the correlation between estimates of student ability obtained from both models. The median reliability for scores in all domains for each of the models was quite similar across groups, with median values ranging from 0.8 to 0.85. There were a few groups where the reliabilities were notably lower (less than 0.6). The inclusion of these data had some influence on the model fit, but there was insufficient evidence based on the reliability to suggest that Model (9.16) should be preferred over Model (9.15). Additionally, the correlation between estimated scores for Models (9.15) and (9.16) in each domain was  $r = 0.999$ , which suggests that there was little added utility in using Model (9.16). We can conclude based on these results that model (9.15) describes the data sufficiently well.

This means that there is a need to specify item-specific, but not person- (or country<sup>2</sup>-) specific, mode effect parameters.

**Table 9.17 Measurement invariance assessment using mode effect models for the PISA field trial data, analysed separately for the domains of financial literacy, maths, reading and science**

Domain	Model	Penalty AIC	AIC	Penalty BIC	BIC	Penalty CAIC	CAIC	Log Penalty	Akaike
Financial literacy	(9.14)	192	253996	1003	254807	1099	254903	0.564498	0.564925
Financial literacy	(9.15)	236	251899	1233	252896	1351	253013	0.559736	0.560260
Financial literacy	(9.16)	248	251744	1295	252792	1419	252916	0.559365	0.559917
Maths	(9.14)	620	1416987	3697	1420064	4007	1420374	0.526304	0.526534
Maths	(9.15)	674	1409948	4019	1413293	4356	1413630	0.523668	0.523919
Maths	(9.16)	714	1409235	4257	1412778	4614	1413135	0.523388	0.523654
Read	(9.14)	818	1770885	4877	1774944	5286	1775353	0.534144	0.534391
Read	(9.15)	990	1760709	5903	1765622	6398	1766117	0.531022	0.531320
Read	(9.16)	1104	1758594	6583	1764073	7135	1764625	0.530349	0.530682
Science	(9.14)	1694	5378045	10100	5386451	10947	5387298	0.586249	0.586433
Science	(9.15)	1984	5361306	11830	5371152	12822	5372144	0.584392	0.584608
Science	(9.16)	2180	5356556	12998	5367374	14088	5368464	0.583852	0.584090

An evaluation of the log-penalty shows that the simple item-independent mode-effect model does not fit as well as the item-specific Model (9.15) and the Model (9.16) with an additional latent variable. Models (9.15) and (9.16) appear to fit the data almost equally well, both accounting for item-specific effects in slightly different ways. Therefore, it can be assumed that a mixture of strong and weak factorial invariance holds and that the computer-based version of the test measures the same construct as the paper-based version. Clearly, the mode effect is not a homogenous shift of difficulties, but rather one that affects some items more than others; a large percentage of items show strong invariance and are not affected in a significant way by mode differences. Further, the results of estimating Model (9.15) for each domain showed that most mode effects on individual tasks were positive, although some were negative. This result shows that a common linear adjustment-based equating method would not be appropriate, and it opens opportunities to optimise the linking between paper- and computer-based assessments by means of item selection, and equality constraints for those items that are least affected by changes in presentation mode.

The distribution of the mode-effect sizes indicated that we can identify a set of items for which strong measurement invariance holds. Those items for which no significant mode effect could be detected formed the basis for linking the computer-based assessment to past PISA cycles, while all trend items can be used, if retained in future studies, to measure the construct due to the invariance properties established in this section.

In summary, the model that balances complexity and model data fit for evaluating and accounting for item mode effects among those considered here was the model that assumes the same parameters for the paper-based assessment as for the computer-based assessment and adjusted the paper-based item difficulty parameters by a differential item-functioning parameter for a subset of items, without the introduction of an additional mode-specific skill. This indicated that *strong measurement invariance can be established for the majority of items* while weak factorial invariance could be assumed for the remaining trend items administered in the computer-based PISA field trial.

It is important to point out that these results indicate that the computer- and paper-based trend items for PISA 2015 can be linked using this approach based on established measurement invariance. The adjustment, if necessary, for a number of items appears to be small compared to the range of difficulty parameters in the trend item set, while the direction of adjustment points to added difficulty.

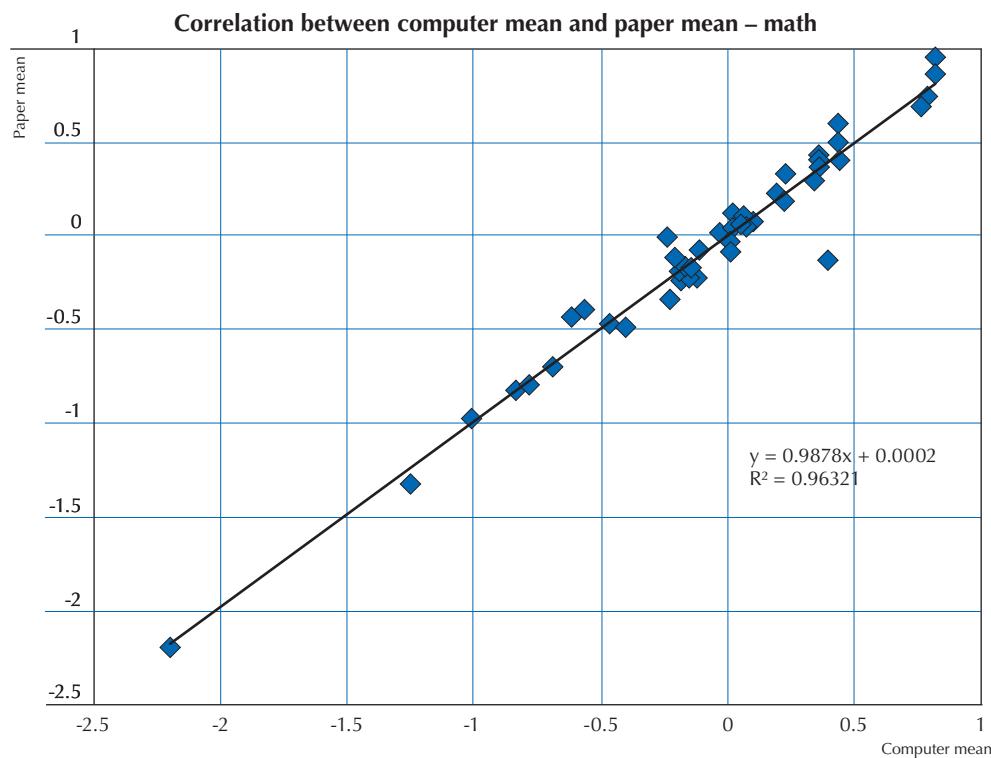
### The impact of mode effects on country means in the field trial

To evaluate the impact of mode effects relative to other variables of interest, country means based on the domain-specific skills obtained from a simplified version of Model (9.16) were split by three variables and compared to one another: gender, mode and a random split of schools within each country. Model (9.16) was simplified for this analysis so that it incorporates scalar invariance for those items that showed little or no mode difficulty differences and assumes metric invariance for the remaining items. There were no country-specific mode effects needed or applied in these analyses. This ensures comparability across countries while accounting for item-specific difficulty differences for a subset of items only, with these differences applied across all countries in the same way. This approach ensured that comparability is maximised, while mode effects that affected different items in different directions were accounted for so that potential effects on scale comparisons were minimised.

The comparisons are illustrated in Figures 9.11 to 9.19 separately for the domains of reading, mathematics and science. These figures show that for each domain, good agreement between country means by assessment mode could be achieved. The largest differences between means were observed based on a random school split, not based on mode. Thus, differences between countries might be due more to differences between students and schools than to differences based on the mode of assessment.

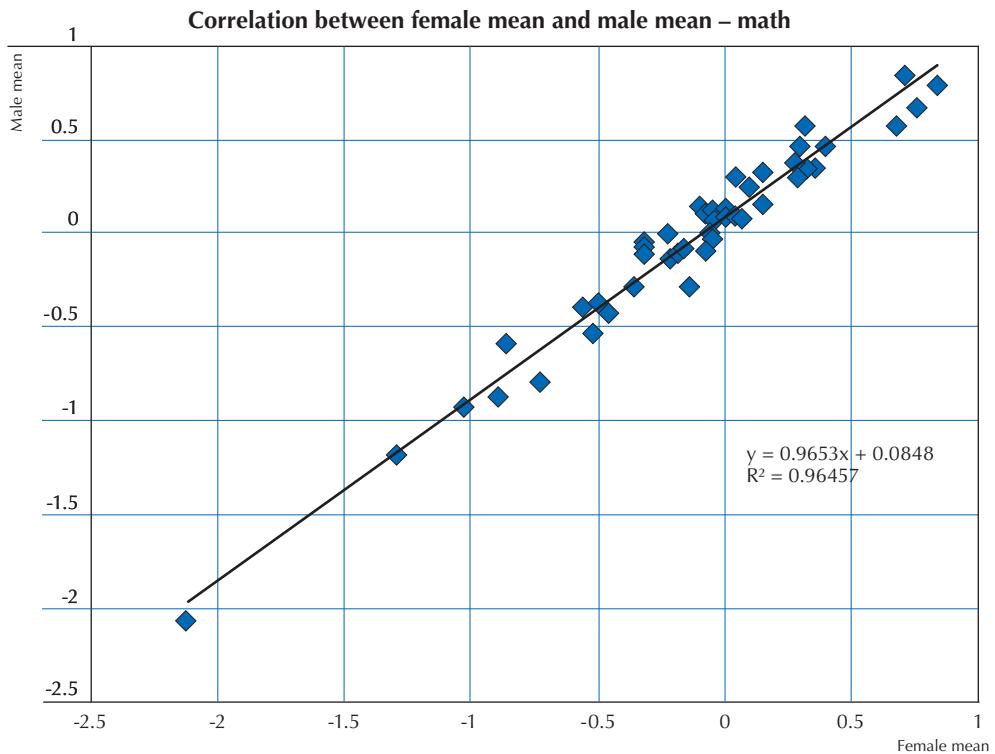
In summary, the differences and variability between gender groups and also the two groups formed by randomly splitting the 25 schools in the field trial were at the same level or larger than the differences obtained by splitting the sample by mode (in other words: mode effects do not seem to be the biggest problem). The apparent mode differences that may be observed if individual countries split their data by mode have to be viewed in the light of these results. Given the sample size of the field trial, differences that one may be tempted to attribute to mode differences are at the same order of magnitude as what could be observed if we split the field trial sample randomly by some other criterion.

■ Figure 9.11 ■  
**Split of country means by assessment mode for mathematics**

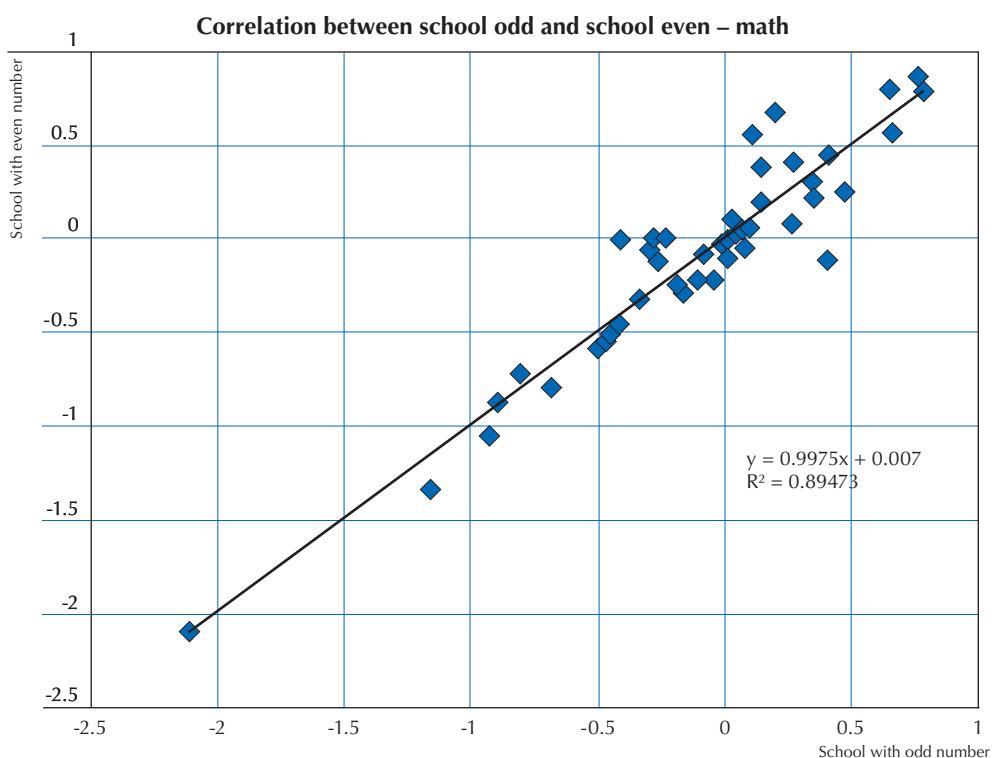




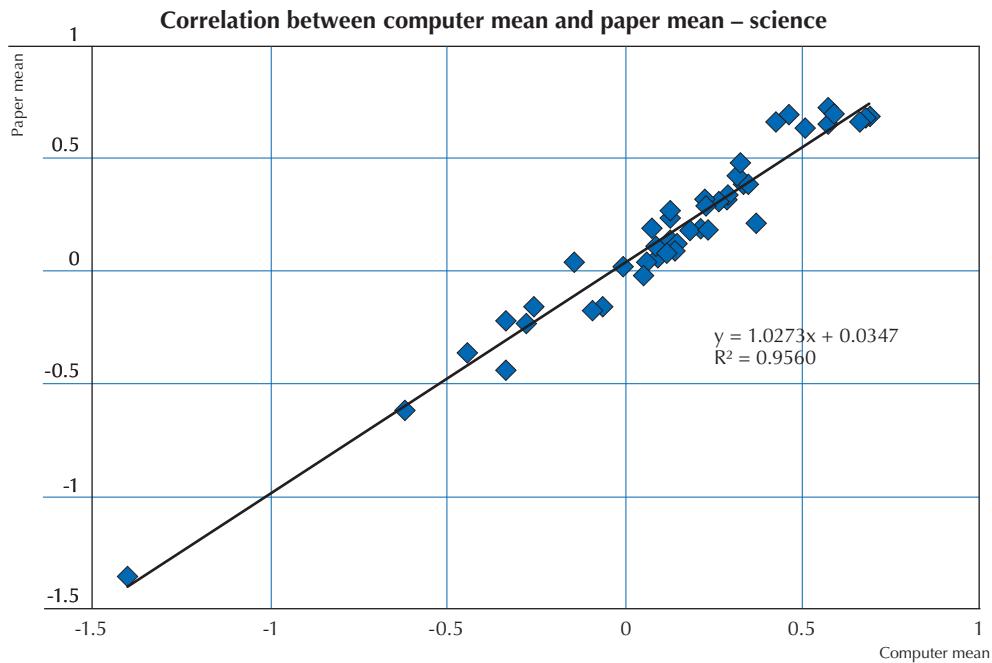
■ Figure 9.12 ■

**Split of country means by gender for mathematics**

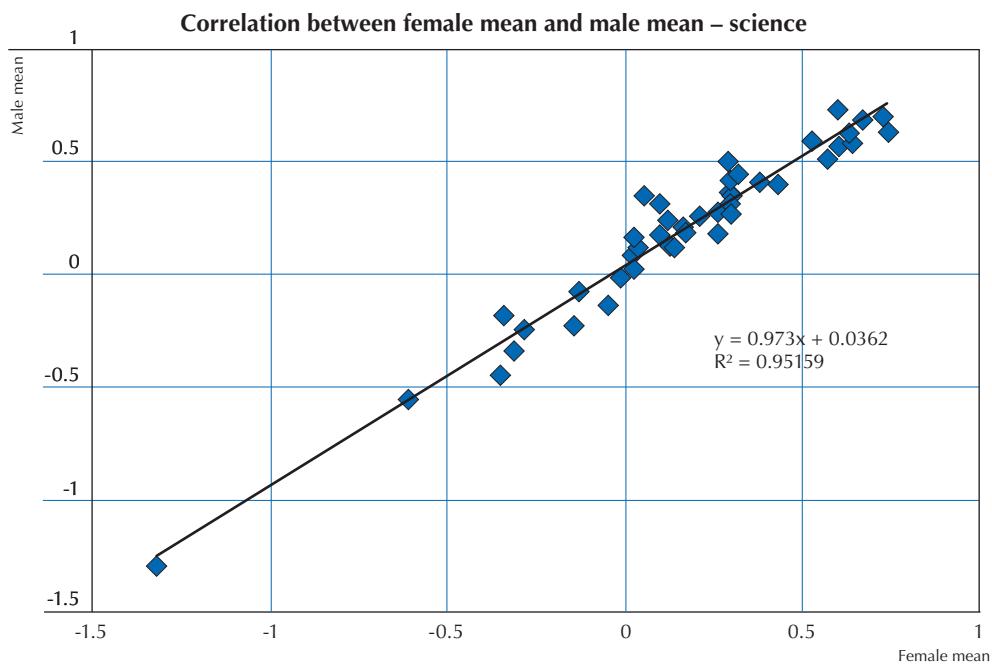
■ Figure 9.13 ■

**Split of country means by random school split for mathematics**

■ Figure 9.14 ■

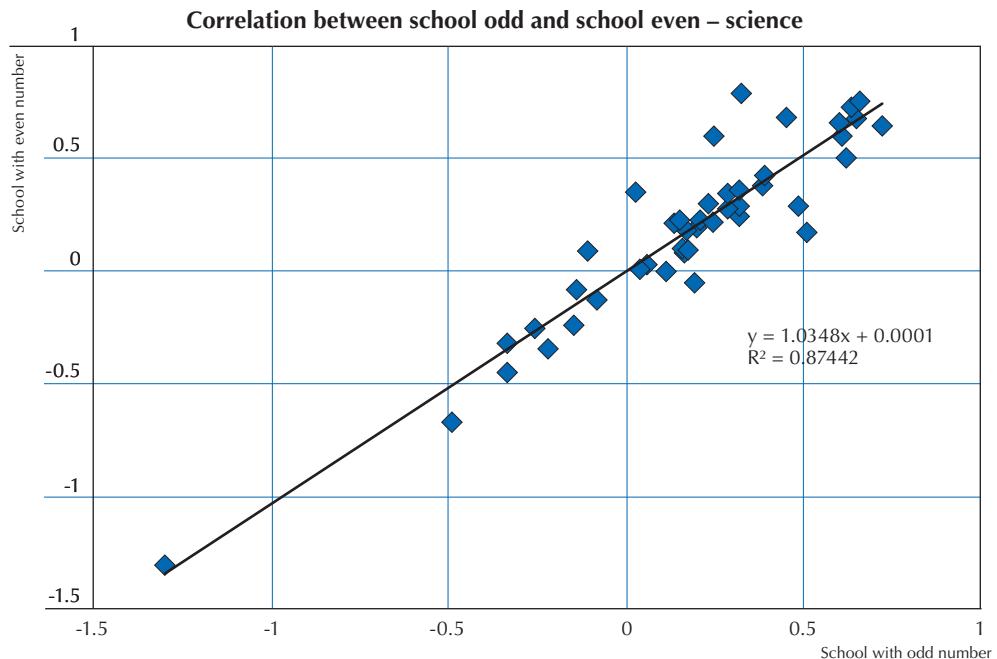
**Split of country means by assessment mode for science**

■ Figure 9.15 ■

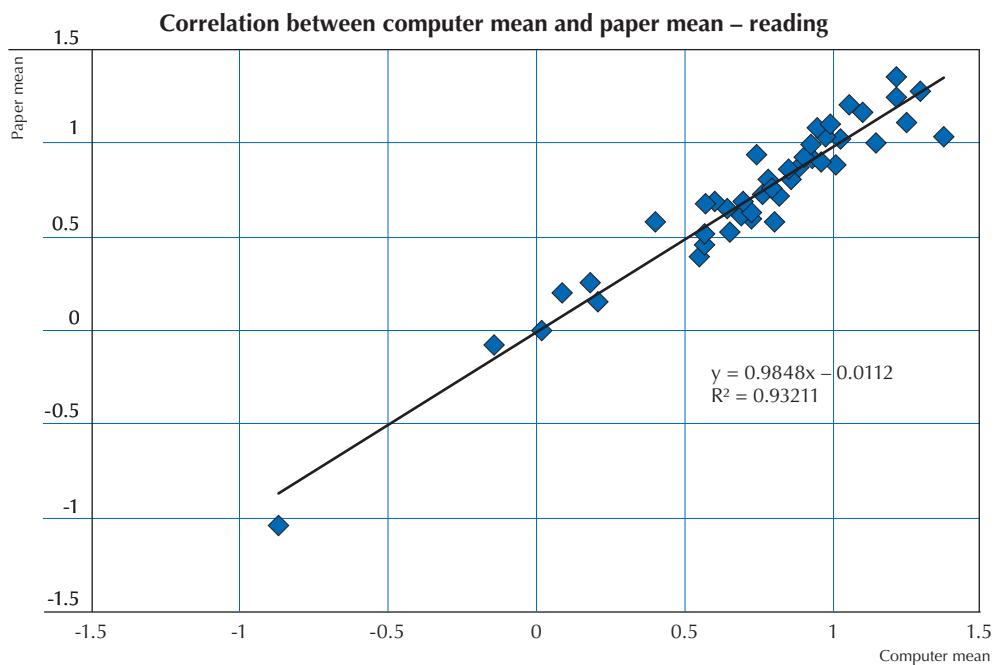
**Split of country means by gender for science**



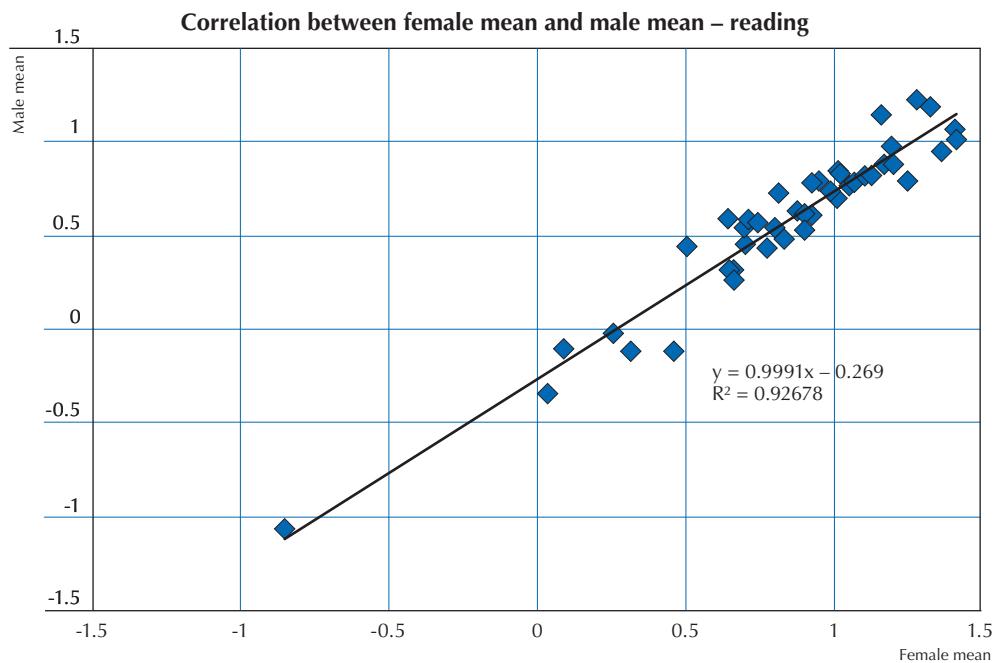
■ Figure 9.16 ■

**Split of country means by random school split for science**


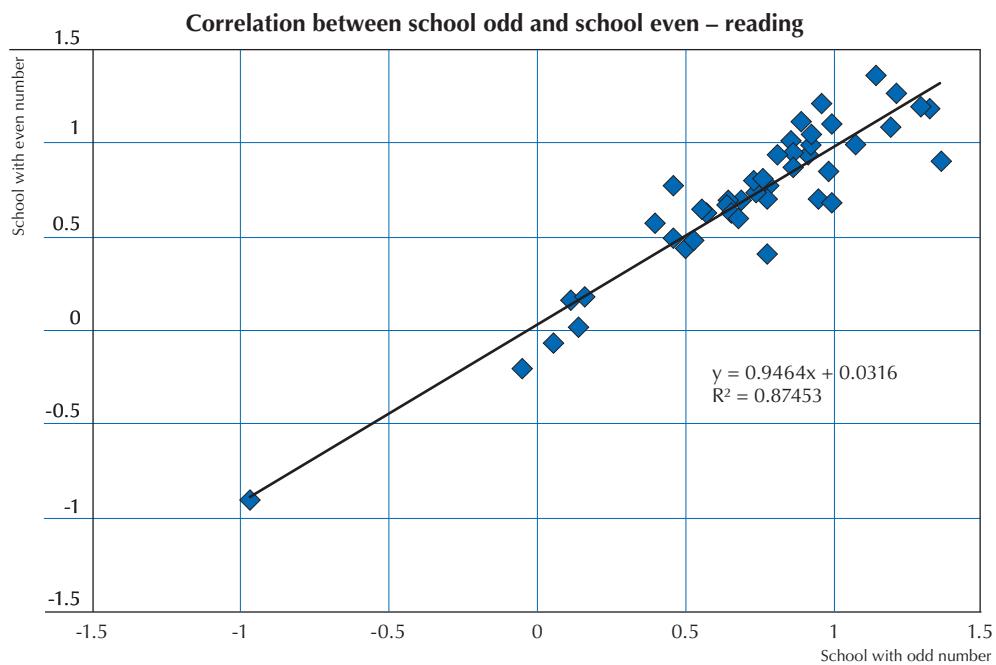
■ Figure 9.17 ■

**Split of country means by assessment mode for reading**


■ Figure 9.18 ■

**Split of country means by gender for reading**

■ Figure 9.19 ■

**Split of country means by random school split for reading**



## Dimensionality and scaling of science trend and new items

### Dimensionality of the science scale

The new science items developed for 2015 are based on a revised assessment framework for this domain. These new items exist in the computer-based assessment mode only because PISA 2015 represents a shift from a paper- to a computer-based survey. In addition to the 85 trend science items from previous PISA rounds, the science domain in the main survey consists of 99 new items resulting in a total of 184 overall. The scales for all PISA content domains have historically been based on the assumption that all underlying constructs are unidimensional. With the revised framework for science it is important to evaluate whether the unidimensionality assumption still holds before new and trend items can be scaled together.

This assumption was tested by comparing a unidimensional model (where new and trend items were assigned to the same unidimensional factor) and a 2-dimensional (multidimensional) confirmatory IRT model (where new and trend items were assigned to two different factors). In addition, a Rasch model for the unidimensional science scale was provided as comparison. All models, the Rasch, the two-parameter logistic /generalised partial credit model and the 2-dimensional (multidimensional) confirmatory IRT model two-parameter logistic/generalised partial credit model were estimated as multiple group models using country-by-language groups. The data used for this analysis came from the subset of computer-based assessment countries that was available at the end of March 2015; please note that due to the potential on the analysis of the PISA 2015 data, this analysis had to be completed prior to analysing the data from all PISA computer-based assessment countries.

Results based on overall model selection criteria show that the unidimensional two-parameter logistic/generalised partial credit model should be preferred over the 2-dimensional model (see Table 9.18). The difference in model fit improvement based on the Gilula and Haberman (1994) log penalty measure is negligible. The two-parameter logistic/ generalised partial credit model reaches 99.91% of the model fit improvement compared to the 2-MIRT model, both in reference to improvement over the independence (baseline) model. Moreover, model-based correlations obtained from the 2-dimensional model show high correlations between the two factors (new and trend items) ranging from 0.83 to 0.96 across the different groups, suggesting there is a single identifiable underlying latent variable. Additionally, the dimension-specific weighted likelihood estimates (WLEs) of student ability are very highly correlated with the unidimensional WLEs. Hence it is reasonable to assume that new and trend science items and scores can be placed on the same unidimensional scale.

**Table 9.18 Model selection criteria for the unidimensional and the two-dimensional IRT models for trend and new science items**

	AIC	BIC	Log penalty	% improvement
Independence	NA	NA	0.6479	0.00%
Rasch model	8021282.185	8024639.114	0.5720	90.88%
2PL/GPCM	7916247.615	7922743.894	0.5645	99.91%
MIRT 2-dimensions	7915262.270	7922400.924	0.5644	100.00%

Note: Log penalty (Gilula and Haberman, 1994) provides the negative expected log likelihood per observation, the % Improvement compares the log-penalties of the models relative to the difference between most restrictive and most general model. The two-parameter logistic/generalised partial credit model reaches 99.91% of the likelihood improvement compared to the 2-dimensional MIRT model, while the Rasch model reaches 90.88%.

### Residual Analysis for Science

As additional evidence in support of the unidimensionality assumption for the science scale, a residual analysis was conducted for the new science items. Due to the nature of the new science items (simulation-based tasks, including different steps for the students to follow) the goal was to investigate possible local dependencies among items. If such dependencies are present, this would pose a threat to the assumption of a unidimensional scale.

First, response residuals were calculated for each item response and correlations among residuals (across respondents) were computed. A principal component analysis using the resulting correlation matrix was then conducted. The principal components analysis was used to evaluate the dimensionality of the scale. Should the first component among residuals be much larger than the second component, an additional latent trait other than the overall ability would be assumed.



Response residuals were computed after the item calibration process in each domain using the *mdltm* software (von Davier, 2005). For dichotomous item responses, response residuals for a person  $v$  with estimated ability  $\hat{\theta}_v$  for each item  $i = 1, \dots, K$  were defined as below:

**9.17**

$$r(x_{iv}) = \frac{x_{iv} - P(X_i = 1 | \hat{\theta}_v)}{\sqrt{P(X_i = 1 | \hat{\theta}_v)[1 - P(X_i = 1 | \hat{\theta}_v)]}}$$

For polytomous item responses, response residuals were calculated using the conditional mean and variance defined below.

**9.18**

$$r(x_{iv}) = \frac{x_{iv} - E(X_i | \hat{\theta}_v)}{\sqrt{V(X_i)}}$$

**9.19**

$$E(X_i^m | \hat{\theta}) = \sum_{x=1}^{\max(X_i)} x^m P(X_i = x | \hat{\theta})$$

**9.20**

$$V(X_i | \hat{\theta}) = E(X_i^2 | \hat{\theta}) - [E(X_i | \hat{\theta})]^2$$

Response residuals were calculated for the 99 new science items using data from a subset of computer-based assessment countries (46 countries). Note again that due to the timeline of PISA 2015, this analysis was completed prior to receiving the data from all PISA countries.

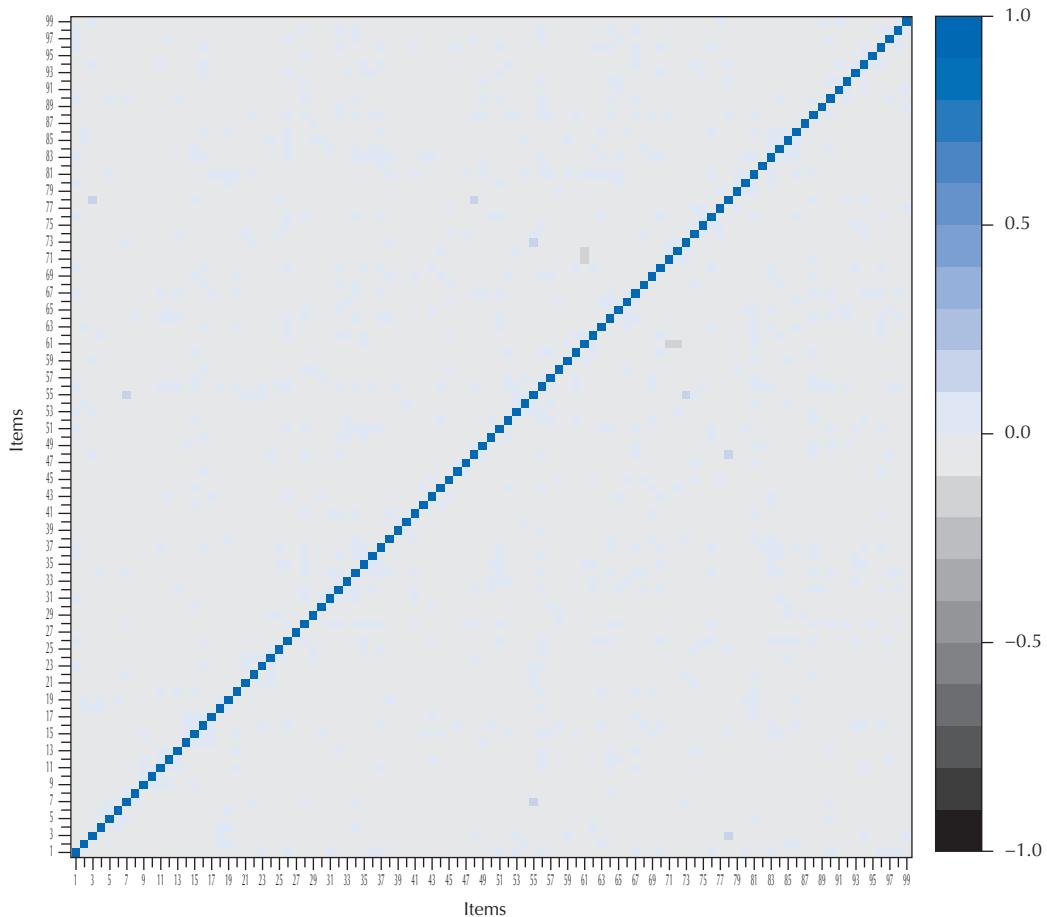
In PISA 2015, no student responded to all of the questions. Given this missing-at-random design, Pearson correlations among items were calculated via pairwise deletion. The visual representations of the correlation matrices were evaluated for remaining dependencies. When a pair of items showed higher correlations, the pattern was checked to determine if it was consistent across countries. Findings from the correlation matrix were interpreted in connection with the item slope parameter estimates and item-total correlations. If an item pair showed highly correlated response residuals and the item slope parameter estimates were high as well for both items, converting these two item scores into a sum score and treating this score as one polytomous item was considered (Rosenbaum, 1988; Wilson and Adams, 1995).

Figure 9.20 shows a heat map plot of the correlations among item level response residuals for the new science items, averaged across countries. Highly-positive correlations between item pairs would be indicated by blue diamonds, highly-negative correlations would be indicated by red diamonds. Since there are none apart from the expected perfect correlation of each residual with itself, this plot suggests that there are no remaining local dependencies among the items after controlling for the latent ability. This pattern was consistently observed across countries. These findings, as well as the results of the principal component analyses, show that there are no local dependencies among the items. Hence, no further treatment (combination or exclusion of items) was needed for new science items.



■ Figure 9.20 ■

### Correlation plot among new science items averaged across countries (46 countries)



#### Final scaling of science in the main survey

After confirming that all science items can be calibrated unidimensionally and without the need to change the scoring of the new simulation-based items, all items were calibrated using a single-scale multiple-group IRT model. No item had to be excluded from the calibration. The IRT scaling was conducted using the 2015 data together with the historical PISA data (2006–2012). The estimation of international/common item parameters and unique item parameters, in case of item misfit, and the treatment of items with identified mode effects followed the procedure described earlier.

The IRT calibration results show very good fit of the international item parameters. The international/common item parameters for both new and trend items were retained for 89.7% of trend items and for 93.3% of the new science items (see Chapter 12 for more information about scaling outcomes).

#### Scaling of reading and mathematics

In the PISA 2015 main survey, the domains reading and mathematics consisted of trend items only. Mathematics comprised 83 trend items in the paper-based assessment (PBA) and 82 equivalent trend items in the computer-based assessment (CBA). Reading consisted of 103 trend items in the PBA and 103 equivalent trend items in the CBA. Both domains were scaled separately using unidimensional multiple-group IRT models (see *The IRT models for scaling above*). The IRT scaling was conducted using the 2015 data together with the historical PISA data (2006–2012). The estimation of international/common item parameters and unique item parameters, in case of item misfit, and the treatment of items with identified mode effects followed the procedure described in the sections *National and international item calibration* and *Handling of item-by-country/language and item-by-mode interactions* earlier in this chapter. One mathematics item had to be excluded from the scaling (see Table 12.1 in Chapter 12); no items were excluded for reading.

The IRT calibration shows very good fit of international/common item parameters. The international parameters were retained in 89% of cases for common item parameters for reading items and in 94.5% of cases for items from the mathematics scale (see Chapter 12 for more information about scaling outcomes). The results illustrate high comparability of the results across different countries and languages, and across different assessment cycles and assessment modes.

## Dimensionality and scaling of collaborative problem solving

### Dimensionality of collaborative problem solving in the field trial

The collaborative problem solving (CPS) scale in the 2015 PISA field trial consisted of 7 units that comprised 188 items. The units are based on simulated conversations with one or more computer-based agents that are designed to provide a virtual collaborative conversation. Students have to choose an optimal sentence from a multiple-choice list to go through the conversation with agents, or choose one or more actions programmed in the unit.

For two of the seven units (unit 101 and unit 105) changes to the scoring of responses were necessary before the data could be used for IRT scaling. Using path analyses, it was found that – due to the nature of the collaborative problem solving items – data from the two mentioned units showed item dependencies in the responses. This was because of different paths that could be taken by students through the simulated chat, resulting in negative residual correlations. Since such dependencies have the potential to introduce bias into the results, the collaborative problem solving chat items exhibiting dependencies were combined into polytomous “composite items” by summing the responses for the different paths students could take. Table 9.19 provides an overview of the combination rules used for these composite items. Given these combinations, the number of items available for the IRT scaling was 164.

**Table 9.19 Combination of collaborative problem solving items of Units 101 and 105 to achieve fair scoring in the PISA 2015 field trial**

New item ID for composite items	Combinations of CPS items
CC101201C	CC101201+CC101202
CC101203C	CC101203+CC101204+CC101205
CC101206C	CC101206+CC101207
CC101301C	CC101301+CC101302+CC101303
CC101304C	CC101304+CC101305
CC101307C	CC101307+CC101308+CC101309A+CC101309B+ CC101310+CC101311+ CC101312A
CC101312BC	CC101312B+CC101313
CC101317C	CC101317+CC101318+CC101319
CC105103C	CC105103+CC105104
CC105105C	CC105105+CC105106+CC105107
CC105201C	CC105201+CC105202
CC105208C	CC105208+CC105209+CC105210
CC105212C	CC105212+CC105213
CC105304C	CC105304+CC105305

### Dimensionality analysis of collaborative problem solving field trial data

The different units were combined into four clusters presented as C1 to C4 in the assessment design. The correlations between the clusters in the Field Trial were generally reasonable, with a range from 0.76 to 0.81 except for those involved with C1. Cluster 1, which contained only a single unit, had lower correlations with the other clusters, ranging from 0.69 to 0.73.

The specific structure of the CPS units and response types, as well as the results from the IRT analysis of the CPS using the unidimensional models, prompted the need to conduct additional analyses (discussed below). However, the unidimensional IRT models showed acceptable fit in terms of item mean deviation and root mean square deviation.



The structure of the CPS units was such that there were a relatively large number of response variables within a unit, while the number of units was small. The contextual coherence of the chat selections that made up these responses followed a common theme within a unit; the conjecture thus could follow that what is measured is more the understanding of what a particular topic requires and might therefore be very specific to each unit.

In order to examine this question, the collaborative problem solving data from the PISA 2015 field trial were analysed using multidimensional IRT models, more specifically with a bifactor model (Holzinger and Swineford, 1937). This model allows an evaluation of whether there is a single source of common variance shared across units, or whether the observed responses are additionally driven by unit-specific response tendencies. In other words, the bifactor model, when compared to a unidimensional model, allows a test of whether unit-specific factors have to be taken into account.

**Table 9.20 Comparison of two-parameter logistic/generalised partial credit models and bifactor model for 164 CPS items**

	Likelihood	A-penalty	AIC	B-penalty	BIC
2PLM/GPCM	-971208	1000	1943417	5652	1948069
Bifactor	-962224	2206	1926653	12468	1936915

The results in Table 9.20 suggest that a bifactor model including a latent variable for each unit fitted the Field Trial data better than the unidimensional two-parameter logistic/generalised partial credit models. The bifactor model indicates that unit response variance was due to unique factors that are not fully measured by a latent variable defined across response variables without looking at their association with a specific content or unit.

It turned out that this result was mainly due to a single unit, presented as C1. As a consequence of these findings, one unit (unit 101) was not included in the PISA 2015 main survey. Additional dimensionality analyses (residual analysis, principal component analysis) were conducted with the main survey data in order to further examine and treat local dependencies of collaborative problem solving items. The next section describes these additional analyses and findings based on the main survey data.

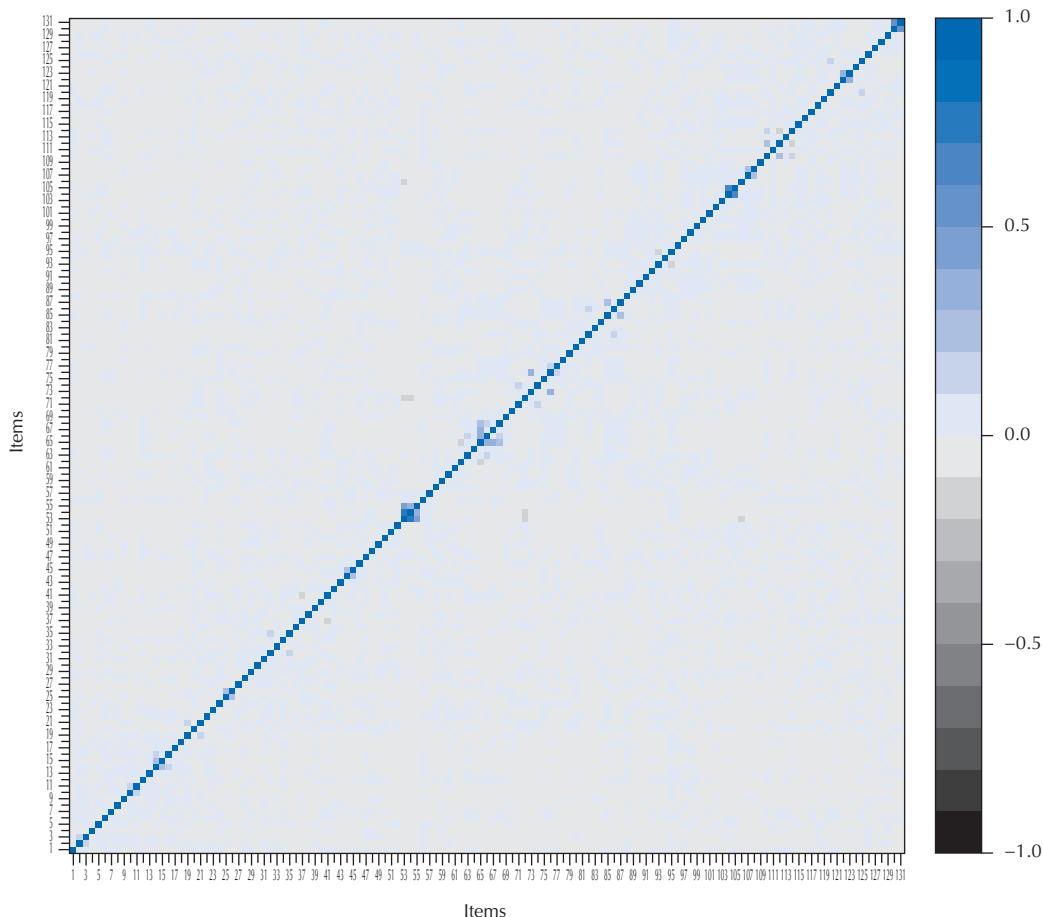
#### **Dimensionality and residual analysis of collaborative problem solving in the main survey**

For the PISA 2015 main survey, 134 items were selected out of the 164 (partly combined) items for the collaborative problem solving domain (unit 101 was not selected). The multidimensional structure of these items was examined residual analyses revealed further dependencies among items that led to further combinations of items into polytomous items (composite items). The residual analyses for CPS followed the same procedure as described earlier for science (*Final scaling of scientific literacy in the main survey*). Item-level response residuals were calculated for each item by respondent interaction for all observed responses, and pairwise correlations among these residuals were computed for the different country samples. Note again that due to the timeline of PISA 2015, this analysis was completed prior to receiving the data from all participating PISA countries. Several pairs of items were identified with highly correlated residuals; the pattern was quite consistent across countries. Figure 9.21 shows the correlations among collaborative problem solving items averaged across countries. Relatively highly correlated item pairs are indicated by blue diamonds and were mainly found near the diagonal line. This indicates that the dependencies (high item-pair correlations of response residuals) were mainly localised and taking place within a few selections. Rather than accounting for these in generalised latent traits measured through all responses in a unit, these localised dependencies were treated by item combinations as described above.

Based on the findings from the residual analyses, additional items were combined into composite items to remove the remaining local dependencies. Table 9.21 shows the combination of these items into composite items. Details about the items included in this rescore can be found in the databases containing country-specific data as well as variable and value labels.

■ Figure 9.21 ■

**Correlation plot among collaborative problem solving items averaged across countries before treating them as composite items (31 countries)**



**Table 9.21 List of composite items based on residual analyses**

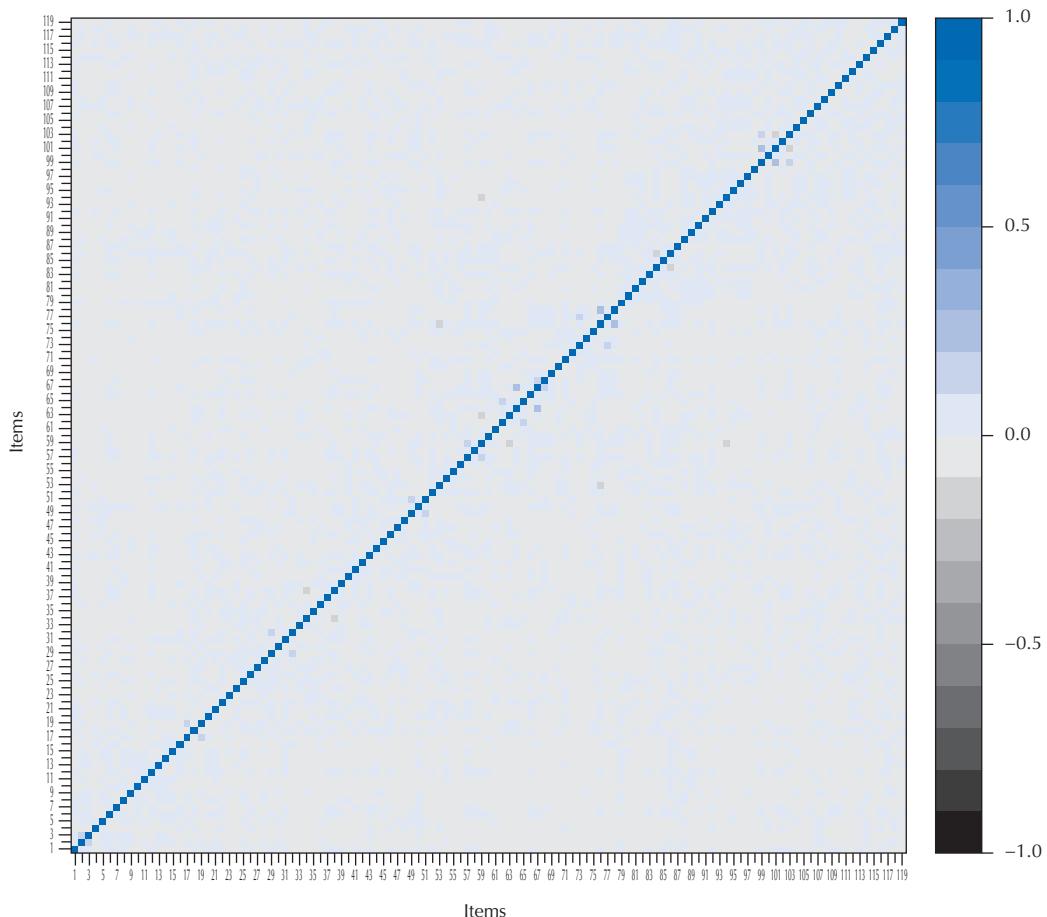
New item ID for composite items	Combinations of collaborative problem solving items
CC104301C	CC104301+CC104302+CC104304
CC106107C	CC106107+CC106108
CC102102C	CC102102+CC102103
CC102209C	CC102209+CC102210+CC102211
CC103108C	CC103108+CC103109+CC103110+CC103111
CC105108C	CC105108+CC105109
CC105203C	CC105203+CC105204
CC105308C	CC105308+CC105309
CC105408C	CC105408+CC105409

After the combination into additional composite items, the number of collaborative problem solving items was reduced to 121 (from the initial set of 134 items) for inclusion in the IRT scaling. In order to evaluate the performance of the composite items, residual analyses were repeated using the 31 countries and 11 additional countries for which data were later received (42 countries in total). Visual representation of the correlation matrix in Figure 9.22 confirmed that remaining local dependencies among items were successfully treated. In contrast to Figure 9.21 that shows several blue diamonds (highly correlated items) near the diagonal line, Figure 9.22 shows no blue diamonds off the diagonal line.



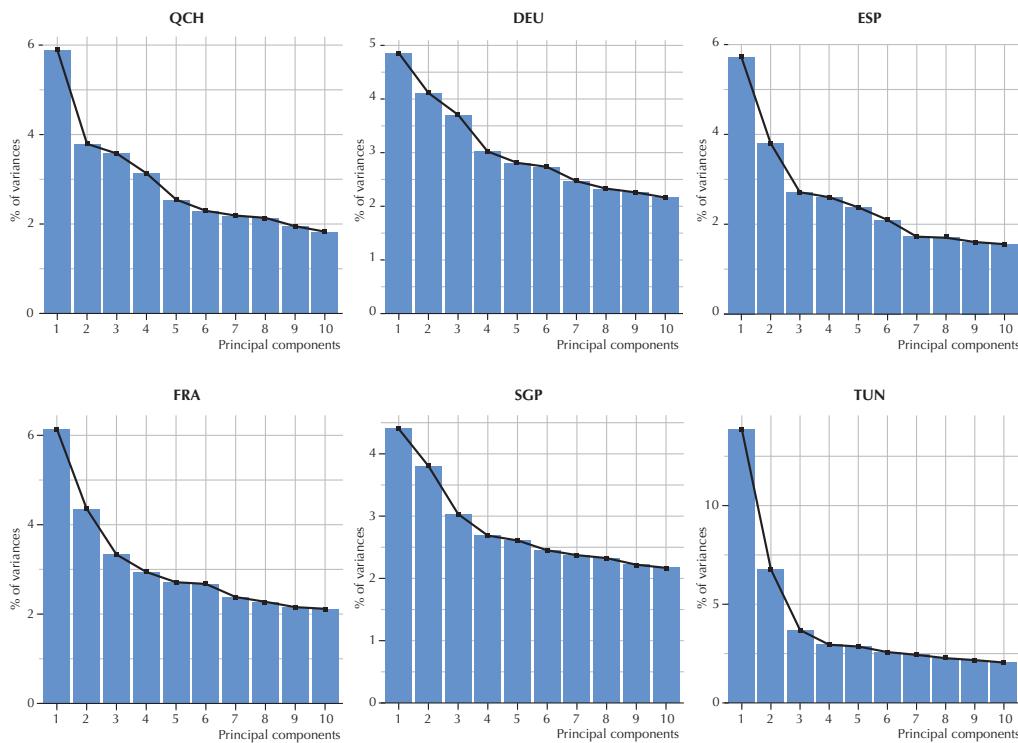
■ Figure 9.22 ■

**Correlation plot among collaborative problem solving items averaged across countries after treating them as composite items (42 countries)**



In addition to the collaborative problem solving residual analysis, a principal components analysis was conducted using the residual correlation matrix. The principal components analysis was used to evaluate the dimensionality of the collaborative problem solving items. Should the eigenvalue of the first principal component extracted from response residuals be large, an additional latent trait other than the overall ability could be assumed. When all items are included as variables, the percentage of variance adds up to 100%. The percentage of variance for the first principal component ranges from 4.4% to 13.9% with a mean of 6.9%. This number can be considered a small amount of common variance. When the percentages of variance for the first 10 principal components are summed up, the value ranges from 26.2% to 41.5%, with a mean of 32.5%, a value that is more typical for a substantial amount to be considered due to a common source of variability of response variables. The small amount of variance of the first, relative to the sum of the variances of the first ten components shows that one cannot justify the assumption of another dimension that may be able to explain statistical dependencies between residuals. In other words, once the ability dimension is accounted for, there is very little common variance among the response residuals.

■ Figure 9.23 ■

**Percentage of variance from principal component analyses (6 example countries)****Operational scaling of the collaborative problem solving main survey data**

After removing all observed local dependencies by combining certain items into polytomous items, the resulting 121 collaborative problem solving items were calibrated using a unidimensional IRT model. Four items had to be excluded from the IRT scaling (due to low item total correlations, too few response in one response category, or technical issues; see Chapter 12), resulting in 117 CPS items on which the item parameter estimations are based. Note that all omitted responses in the CPS domain were scored as not reached (missing) due to differences in the administration of this domain. Omissions in reading, mathematics and science may be the result of intentional skipping of items, as students have the ability to move to the next item without interacting with the current one. In collaborative problem solving, however, students must make a sequence of successive choices and cannot skip forward to avoid a choice. Thus, unobserved responses in CPS items are a result of students taking different paths while working on an item, meaning some paths are not taken. Therefore, unobserved responses do not reflect student skill and need to be treated as not administered. The estimation of international/common item parameters and unique item parameters, in case of item misfit, followed the procedure described in the sections *National and international item calibration* and *Handling of item-by-country/language and item-by-mode interactions* earlier in this chapter.

The IRT calibration shows good fit of the international/common item parameters. International parameters were retained in 95% of the item parameters (see Chapter 12 for more information about scaling outcomes) and, thus, a high comparability of the scale across different countries and languages.

**Scaling of financial literacy**

In PISA 2015, financial literacy had a data collection design that provides stronger connections to data collected in other domains, compared to the PISA 2012 design. That is, every student who took financial literacy also took reading, mathematics, or both, in addition to science. Therefore, PISA 2015 provides a better estimate of the covariance between the core domains and financial literacy. However, because not every country took financial literacy in PISA 2015, there are only a few countries that have data available in both years. As such, the 2015 main survey calibration required data from PISA 2012 as well as the 2015 field trial. This approach provides a sound link for PISA 2015 because, in the 2015 field trial data, a larger group of countries took both the computer- and paper-based assessments (for the mode-



effect study). This is also important since the 2015 administration of financial literacy is based on data collection for a subset of students in a second (afternoon) testing session. All available financial literacy data (2012 main survey, 2015 field trial, and 2015 main survey) were combined for the IRT scaling using a multiple-group IRT model based on an equivalent-groups (for the field trial samples) design for the linking. This particular linking method provides a sound link and is robust against changes in the percent correct observed in the 2015 main survey; the inclusion of the field trial data allows the assumption of equivalent groups since students were randomly assigned in the field trial to paper- versus computer-based assessments.

The equivalent groups design is a method of linking that is common in test equating. While it provides a consistent linking approach, it does not provide information on which items are directly comparable. Neither does it require or assume that the items be invariant across assessment modes, since the comparability is established based on the premise that the distribution of student ability is equivalent across groups. The link in financial literacy is established through common populations, while for the other scales (reading, mathematics and science) it was possible to link across modes and assessment cycles using common items.

In the PISA 2015 main survey, the financial literacy domain consists of 43 trend items. No items were excluded from the scaling. The estimation of international or common item parameters and unique item parameters, in case of item misfit, and the treatment of mode effects followed the procedure described in earlier sections.

The IRT calibration shows a very good fit of the international/common item parameters. The scaling was able to retain common/international item parameters for 92.9% of the items (while for 7.1% of the items unique item parameters had to be estimated) and, thus, a high comparability of the scale across different countries and languages (see Chapter 12 for more information about scaling outcomes).

### **Developing common scales for the purpose of trends**

The new modelling approach in PISA 2015 using a hybrid model (the combined Rasch /partial credit model and two-parameter logistic/generalised partial credit model) necessitated a reanalysis of data from prior cycles (2000-2012) with the aim of studying the effect of the more general model applied over multiple cycles on stabilizing the trend measure and to ensure its quality. With the introduction of computer-based assessments as the main mode of assessment in PISA 2015, there was concern that the mode might influence item parameter estimates for the linking items. Moreover, some linking items might not work equally well for all of the populations assessed in PISA 2015. Using these items reduces the comparability of the trend measure; hence, there may be a need to exclude them from the main survey item pool. However, given the new scaling approach for PISA 2015, it might be possible to retain a larger share of these items, since the model used is more flexible and contains the previous scaling approach as a special case.

Results from prior analyses (PISA 2000-2012) were replicated and then re-examined using the hybrid Rasch/partial credit model and two-parameter logistic/generalised partial credit model. The reanalysis produced a common parameter for each of the previously used items in the databases from PISA 2000 to 2012. These parameters were treated as fixed parameters for the PISA 2015 field trial scaling. This was done to establish a stable link between the field trial items and the international scale based on past frameworks of each domain. Parameter constraints for various items were released in subsequent rounds in case of item misfit. The common item parameters in the field trial generally fit well; thus, the same item parameter can be assumed over cycles for a large number of trend items.

The overall item fit for each domain was very good, with small numbers of items misfitting for reading (2.5%), mathematics (1.8%), and science (3.9%). Financial literacy showed the highest percentage of misfit (4.1%). Note that item misfit was defined for root mean square deviation values larger than 0.2 in the field trial then later in the main survey analysis. All of the main scales showed sufficient IRT-based (marginal) reliabilities (Sireci, Thissen and Wainer, 1991; Wainer, Bradlow and Wang, 2007, 76) with 0.83 for reading, 0.81 for mathematics, 0.80 for science (based on trend and new items), and 0.85 for financial literacy. These results illustrate the quality of trend measure across different assessment cycles (2015 data versus 2000-2012), different assessment modes (paper- versus computer-based assessments), and even across different countries and languages as the multi-cycle scaling with common item parameters assures the equivalence of inferences of trend assessment.

In the PISA 2015 main survey a comprehensive rescaling was carried out including the 2015 main data. This was done to ensure that the main survey data equally contributed to the estimation of item parameters, while establishing the link to past PISA rounds by including previous cycles. Instead of fixing the item parameters for trend items

obtained from past (historical) data to the 2015 data, item parameters were estimated based on all available data from 2006 through 2015. The historic data were only included back to 2006 because this was the last cycle when science was the major domain, and because there were very few items left in the 2015 round that dated back to the early (2000 and 2003) rounds of PISA. This approach ensured that domains tested in 2015 with a new design that improved minor domain coverage and broadened the assessment of the revised framework were contributing to the estimates that established the common scale linking across prior PISA cycles. The IRT calibration for each domain showed good fit of the items to the international/common item parameters. The results also showed high comparability in the item parameters across different countries and languages, and across different assessment cycles and assessment modes.

### **Rescaling PISA 2000-2012**

The PISA 2015 field trial and main survey design were premised on the availability of a quality set of the linking items across the previous PISA cycles. These designs incorporated all previously used trend items from all previous cycles in the field trial so that the best possible link could be established.

This increase in scope also required that prior analyses be revisited because the integration of all previously used trend items required a full re-estimation of the scaling model on which prior PISA cycles were based. There is strong evidence in favour of a joint model for linking the cycles across multiple populations (von Davier and von Davier, 2007; Mazzeo and von Davier, 2008, 2014). This also allows different trend clusters containing items sets not previously used in a single assessment to be linked together within a comprehensive modelling approach.

PISA has collected data in representative samples of 15-year-old students around the world every three years since 2000. In each of the first five cycles (2000, 2003, 2006, 2009 and 2012), both OECD and partner countries participated, resulting in almost 300 cohorts defined by assessment year and country. Many of the OECD countries, as well as a substantial number of partner countries, had participated in each of the first five PISA cycles and continued to do so in 2015.

In work leading up to the 2015 main survey analysis, an effort to utilise all available evidence on item functioning and scale coverage of the task material used in PISA was made. ETS compiled a database that merged all five cycles and all countries. This yielded a file that contains roughly 2 million student records. ETS utilised a multiple group IRT model approach to link all items, by domain, across all PISA cycles by country combinations (Bock and Zimowski, 1997; Mazzeo and von Davier, 2014, 2008; von Davier and von Davier, 2007; von Davier and Yamamoto, 2004; Weeks, von Davier, and Yamamoto, 2014; Yamamoto and Mazzeo, 1992).

Several analytical steps were performed. More specifically, in order to find the best fitting model, different and increasingly complex IRT models were specified and estimated; model-data fit was compared using both AIC and BIC as well as measures of item fit. The analyses were carried out separately for each of the main PISA domains of reading, mathematics and science.

In a first step, the model used in the operational reporting of PISA 2000-2012 was recreated in order to ensure that the results obtained in the previous analyses could be replicated. Previous cycles of PISA utilised the mixed-coefficients multinomial logit model (MCMLM; Adams, Wilson, and Wu, 1997), which is a generalisation of the Rasch model (Rasch, 1960) that allows for category weights, multiple populations, and predictors of ability as well as polytomous response data. This was followed by an approach that utilised model-data fit indicators to relax model assumptions of the Rasch model where needed. More specifically, the Rasch model assumption of equal slopes was relaxed if it was found that the item discrimination was markedly different in the group of countries by cycles in which an item was used. ETS compared this analysis with an estimation of the two-parameter logistic /generalised partial credit model (Birnbaum, 1968; Muraki, 1992) for multiple populations (von Davier and Yamamoto, 2004).

This initial analytic step allowed the estimation of slope parameters for those items that were found to discriminate more (or less) well than the items that follow the Rasch model. In the next step, model assumptions were relaxed further. Given that international assessments are translated into multiple target languages, item-by-country interactions are a potential threat to validity. As such, some items in some countries may function differently from how the item generally functions in the majority of countries. For this reason, we added an analysis step that investigates item-by-country (by cycle) interactions in order to catch cases in which an item deviates substantively in one or the other cycles of PISA. This approach follows best practices (Glas and Jehangir, 2014; Glas and Verhelst, 1995; Oliveri and von Davier, 2011, 2014; Yamamoto, 1998). All analyses were carried out using the software *mdltm* (von Davier, 2005). The next



section describes the results of the rescaling with the Rasch model, followed by a description of the model that combines features of the Rasch model and the two-parameter logistic/generalised partial credit model and the model for country-by-item interactions.

### **Results for the Rasch model**

In this subsection, we examine the comparability of rescaled and reported results from previous analyses. We initially fit the data with the Rasch model since it has been the operational model used for reporting PISA results by cycle for the past five assessments. The results from our reanalysis of the data were compared with published results available online. Note that for our analysis, we obtained item parameters and country means by estimating one model across all cycles and all participating countries. This approach differs from the operational approach used in past cycles in that it incorporates all data into the item calibration in order to link the results across cycles. The operational approach, on the other hand, uses only the mean of trend items in two adjacent cycles to find transformation constants in order to put the new scaling results on the old scale. If the fit of the model is perfect (i.e. if item parameters stay the same over cycles), and if the item functions can indeed be fitted by the Rasch model, both methods should produce identical results. In this case, however, the use of all cycles in a single comprehensive estimation of the Rasch model should lead to the most accurate item parameters possible, given the data at hand.

The comparison was carried out using two independent rescaling approaches. In contrast to the operational approach implemented by the contractor responsible for the 2000-2012 cycles, we did not use a random selection of 500 cases per country. Instead we used all data from every country participating in these five PISA cycles. The re-estimation of parameters was conducted either per assessment cycle using the ConQuest software (Wu, Adams and Wilson, 1997) or using all data from all five PISA cycles in a concurrent calibration using *mdltm* (von Davier 2005). The replication effort was done to ensure that we could recover the previously estimated item parameters.

In summary, the reproduction of the original reporting scale was fully successful under both estimated approaches. The correlations between country means as reported by PISA and those reproduced by calibrating all available data in a comprehensive scaling was above 0.998 and, in many cases, especially for the *mdltm* calibrations that used all available data across cycles (0.999 and above). This suggests that there were no issues with the data used to estimate the item parameters. However, the estimation of a comprehensive model using data from all cycles leads to the most consistent item parameter estimates and a scale that is linked in the most rigorous way (see also Chapter 2) across all available PISA cycles.

### **Results for the hybrid ‘partial Rasch, partial two-parameter logistic/generalised partial credit’ model**

Given that we were able to replicate the Rasch model results, we moved on to an approach that combined features of the Rasch model and more general IRT models. Among these models are the two-parameter logistic/generalised partial credit model, which estimates a slope parameter for all items, a hybrid combination of the Rasch model and two-parameter logistic/generalised partial credit model that estimates slope parameters only for items that do not fit the Rasch model, and a model that additionally accounts for item-by-country interactions (IBCI) and estimates unique item parameters for countries and/or country-groups for items that cannot be fitted well using a common international parameter (Glas and Jehangir, 2014; Glas and Verhelst, 1995; Oliveri and von Davier, 2014, 2011; Yamamoto, 1998; Yamamoto and Mazzeo, 1992). Note that all model extensions are exponential family models, and that the operational model, the Rasch model, used in the first five rounds of PISA, is a special case of the extended approach. If the Rasch model indeed fits the data, the extended model will just reflect that, namely by fitting the data with something that very closely resembles the fit of the Rasch model. However, if the extended approach statistically fits the data substantially better than the approach used in previous rounds, this will be visible in model selection criteria.

This hybrid combination of item functions from either the Rasch model or the two-parameter logistic/generalised partial credit model allowed for fitting of a wider range of items compared to using the Rasch model alone. In contrast to the two-parameter logistic /generalised partial credit model being applied to all items, we were able to retain a number of slope parameters that are fixed across items, and hence were able to provide a model that makes the same assumption (an equal slope across items) as past PISA cycles for a subset of items. Table 9.22 gives an overview of the number of items that were retained as “Rasch” items using a common slope parameter of 1.0 in the hybrid model (Rasch/ two-parameter logistic /generalised partial credit model) accounting for IBCI (hybrid/IBCI model).

**Table 9.22 Number of Rasch model items retained in the hybrid/IBCI model**

	Total number of items	Rasch # retained	Rasch % retained
Mathematics	179	77	43%
Reading	223	42	19%
Science	133	19	14%
Financial literacy	40	15	38%

Table 9.23 summarises the improvement in model fit for the domains of reading, mathematics and science. The table shows the results for the Rasch /partial credit model, the two-parameter logistic /generalised partial credit model, and the “hybrid” model (Rasch/two-parameter logistic/generalised partial credit model), with one set of item parameters for all countries, and a model that accounts for IBCI by releasing some country-specific parameters. These results are based on all cycles from 2000-2012 combined for the three domains. In each domain, the IBCI model fits best (as characterised by the BIC), followed by the two-parameter logistic /generalised partial credit model, the hybrid model, and the Rasch /partial credit model. This can also be seen in the concomitant decrease in the number of items-by-country-by-cycle with root mean square deviation values greater than 0.15. Approximately 3% of the items in mathematics, 7% of the items in reading and 6% of the items in science did not fit the Rasch model in one or more countries. On the other hand, around 1% of the items exhibit misfit in reading for the IBCI model and less than 0.1% of the items exhibit misfit in mathematics and science under the hybrid/IBCI model. For all subsequent analyses, the item parameter estimates from the hybrid/IBCI model were used.

**Table 9.23 Changes in model fit summary**

		Rasch/PCM	2PLM/GPCM	Hybrid	IBCI
Maths	# of item-country-cycle deviations BIC	549 26400730	397 26118134	415 26175012	4 25946516
Reading	# of item-country-cycle deviations BIC	1233 30968125	960 30675531	962 30691983	250 30472304
Science	# of item-country-cycle deviations BIC	921 29908518	717 29585732	708 29591677	8 29302806

Total item-country-cycle values: maths = 15,795, reading = 18,603, science = 16,223

Deviations defined as RMSD values > 0.15

### **Fit of the Rasch Model and two-parameter logistic model for new science and collaborative problem solving items in the field trial**

After examining the fit of the new modelling approach developed for PISA 2015 to data from past PISA cycles (2000-2012), described in the sections above, the fit of the Rasch/partial credit model versus the two-parameter logistic/generalised partial credit model was tested for new science and collaborative problem solving items using data from the 2015 field trial (note that this comparison was done in the field trial in preparation for the main study; hence, no similar comparison was needed in the main study). The aim was to investigate whether the two-parameter logistic / generalised partial credit model shows a better fit, as would be expected.

While the item parameters for trend items in the field trial were fixed to those obtained from the reanalysis of previous PISA cycles (2000-2012), the new science and CPS items had to be scaled based solely on the field trial data. For these new scales, both a multigroup Rasch/partial credit model was estimated as well as a multigroup two-parameter logistic /generalised partial credit model. The concurrent calibration (multiple-group IRT model) was used to evaluate whether items were working in the same way across country-by-language groups or if there were item-by-group interactions. Both model approaches were compared (see Table 9.24) and it was found that the two-parameter logistic /generalised partial credit model showed better overall model fit than the Rasch/partial credit model. The item selection for the main survey was based on the two-parameter logistic /generalised partial credit model due to the improved model fit and because more information about each single item was provided.

**Table 9.24 Comparison of the Rasch/ partial credit model and the two-parameter logistic /generalised partial credit model for new items in the PISA 2015 field trial**

	Likelihood	A-penalty	AIC	B-penalty	BIC
CPS					
RM/PCM	-985477.57	686	1971641.15	3877.09	1974832.24
2PLM/GPCM	-971208.69	994	1943411.38	5617.83	1948035.21
Science					
RM/PCM	-2215483.30	1266	4432232.60	7406.46	4438373.06
2PLM/GPCM	-2192778.99	1698	4387255.97	9933.78	4395491.75



### **Linking PISA 2015 to previous PISA cycles**

The goals of the PISA 2015 linking design centred on linking different test forms and assessments modes (paper- and computer-based) within the PISA 2015 cycle for comparability across countries and linking previous PISA cycles to PISA 2015 for comparability across assessment cycles and trend reporting.

To obtain comparable test results across the years in each domain, it was important that all items in a given domain were calibrated on one common scale. To establish a common scale, the items had to be linked together across test forms (subset of items), assessment modes (paper- and computer-based), and PISA cycles. This was achieved by using common sets of items in the different booklets and assessment modes. Moreover, the PISA 2015 linking design included items from the previous studies and links all PISA cycles (2000 through 2015). Note that for the scaling in the 2015 main survey, combined PISA data sets from 2006–2015 were used for parameter estimation. The new part of the science scale and collaborative problem solving as a new domain comprised only computer-based items (due to the nature of the items); because collaborative problem solving is a new domain, there are no linking items. Financial literacy, as an optional domain, was only linked back to 2012 (the first time financial literacy was assessed) in the 2015 main survey scaling.

In summary, the computer-based assessments included all domains and all linking/trend items (providing a link between paper- and computer-based testing and between the current and previous PISA cycles) as well as new items for science and collaborative problem solving. The computer-based assessments comprised the following item sets:

- reading, mathematics and financial literacy: intact clusters of paper-based items from previous cycles, reauthored for computer delivery
- science: intact clusters of paper-based items from previous cycles, reauthored for computer delivery, plus new items developed for computer delivery only
- collaborative problem solving: new items developed for computer delivery only.

Thus, all trend items were administered in both the paper- and computer-based assessments as well as in different test forms (across the different assessment modes). Within both assessment modes, all items were linked together in a booklet design, which relates to trend items in the paper-based assessments and the trend and new items in the computer-based assessments. The mode effect study allowed identification of scalar and metric invariant items across computer- and paper-based testing and thus allowed linking across modes. The inclusion of all non-released items in the new assessment design strengthened the construct coverage of the major and minor assessment domains and allowed linking the new science domain against all trend material dating back to the last major domain round in science, assessed in 2006.

The improved linking design established in 2015 (see Chapter 2) made it possible to calibrate all trend and new items answered by different students in different test forms and assessment modes on one common scale for each domain. This was done within the item calibration utilizing the approach described in the sections *The IRT models for scaling, national and international item calibration* and *Handling of item-by-country/language and item-by-mode interactions*.

To place the PISA 2015 results and the historic PISA results from cycles 2012 to 2006 on the same scale, a concurrent item calibration was used. This linking approach is different from the mean/mean IRT linking approach used in prior PISA cycles. For trend items that did not show mode effects, item difficulty, and slope parameters in the main survey were constrained to have the same parameters as the corresponding paper-based items and items found in the historical data, establishing scalar invariance for a majority of items in each domain. For the remaining items, metric invariance was established so that a common slope parameter is shared across cycles and across modes in 2015. This approach created a scale that allowed for the comparison of PISA 2015 main survey and historic PISA results.

For financial literacy, a slightly different approach was taken by linking the 2015 main survey data not only to the data from 2012 but from the 2015 field trial. The reason is that not every country took financial literacy in 2015, and only a few countries took the assessment in both cycles (2012 and 2015). Moreover, the administration of financial literacy in 2015 was based on the data collection from a subset of students in a second (afternoon) testing session. Consequently, linking through the 2015 field trial data, where a larger group of countries took both the computer- and paper-based versions, provides a more defensible scale.

More detailed information about the test design for PISA 2015 can be found in Chapter 2 and more information about the linking and IRT scaling in general and for each domain is given in the relevant sections of *Application of IRT and population models to PISA* above.

### **Linking error in PISA 2015**

PISA accounts for student sampling error, measurement error of ability estimates and linking error. An evaluation of the magnitude of linking error can be accomplished by considering differences between reported country results from previous PISA cycles and the transformed results from the rescaling. Recall that prior PISA rounds used a separate item calibration for each cycle. That is, the same items, if repeatedly used in 2000, 2003, ..., 2012 received slightly different statistical quantities as estimates of their difficulty, especially because they would be tested together with other sets of items, or as part of a smaller (minor domain) or larger (major domain) set of items.

This variability over time and different PISA assessment designs (minor/major, etc.), and also the fact that we do not “know” the difficulty of items exactly, introduces a source of uncertainty in the results. It becomes apparent as soon as there are multiple samples that were collected successively, as the item difficulty parameter estimates tend to be (slightly) different every time new data is collected. This, in turn has an effect on the results reported to countries, and it is (and was in previous cycles) quantified in the linking error. This linking error is a part of the variability of country means that is due to the tests not being exactly the same and having different samples of students in the estimation of item parameters.

The extended analytical approach used in 2015 allows us to revisit the linking error and to reduce it when moving forward with the new design, which reduces construct coverage differences between minor and major domains and with the concurrent calibration used in the IRT scaling.

In summary, the uncertainty due to linking can result from changes in the assessment design or the scaling procedure used, such as:

1. different calibration samples used to estimate parameters in different cycles
2. the inclusion of items that are unique to each cycle in addition to common items
3. changes in the cluster position within the assessment (PISA 2000 was an unbalanced design; later designs balanced cluster positions)
4. changes in the model used for scaling
5. the particular set of trend items that are common to all assessment cycles of interest, and which can be seen as one among an infinite set of possible trend items.

In PISA, it is important to note that the composition of the assessment in any two cycles are different due to Major-minor-minor (M-m-m) domain changes, cluster changes and units released and recombined, framework changes, assessment mode changes, and test design changes. Although the reporting model remains a unidimensional IRT model, which fits quite well, trend items are modelled based on data collected in different contexts (M-m-m or mode, etc.). Thus, estimating linking error for trend measures is a key tool to account for cycle-to-cycle differences. Note again that linking error estimates quantify the uncertainty about the link of a scale value compared between two assessment cycles.

In practice, not all of the sources of uncertainty around scale comparability were quantified or could be accounted for in past PISA cycles (2000-2012). The linking error estimated in past PISA cycles accounted only for differences across trend items observed for the re-estimated difficulty parameter from one cycle to the next. This approach of linking scales by “separate calibrations” includes the following steps. First, calibrations of data from assessment cycle one (Y1) and assessment cycle two (Y2) were run separately with trend items and items unique to each assessment cycles; this produces two sets of trend item parameter estimates (one set for each cycle). The mean of Y2 trend item difficulties was then transformed to the mean of Y1 trend items, in order to link the scales. This mean-only transformation is only valid for the Rasch model, if it is indeed fitting the data. Because the same “shift” parameter is added to all participating countries in order to equate results to previous assessments, any uncertainty that is introduced through this shift is common to all students and all countries. This is a form of mean/mean IRT linking, a method that operates on independently estimated item parameters. This method was applied in past PISA cycles (before 2015), and it relied directly on parameter invariance assumptions in the trend item set comparing estimates from two separate calibrations across adjacent PISA cycles. In this approach, the variance of differences between trend item estimates from the Y1 and Y2 calibration was used to characterise linking error; it can be written as:

**9.21**

$$LE_{<15} = \frac{1}{k} \sum_{i=1}^k (\hat{b}_{Y1,i} - \hat{b}_{Y2,i})^2$$



When we assume that each item parameter estimate can be written as the sum of the true parameter and an error term unique to the cycle, we can write for both cycles:

**9.22**

$$\hat{b}_{Y1,i} = b_{true,i} + \hat{u}_{Y1,i}$$

and

**9.23**

$$\hat{b}_{Y2,i} = b_{true,i} + \hat{u}_{Y2,i}$$

Assuming that the parameter estimates are unbiased yields that both error terms are vanishing in expectation. A final step combines the two separate cycle calibration errors, that is:

**9.24**

$$\hat{e}_{Y1,Y2,i} = \hat{u}_{Y1,i} - \hat{u}_{Y2,i}$$

Then, the pre-2015 linking error in (9.21) can be written as:

**9.25**

$$LE_{Y1,Y2} = \frac{1}{k} \sum_{i=1}^k \left( \hat{e}_{Y1,Y2,i} \right)^2 = V(\hat{e}_{Y1,Y2})$$

The expression in (9.25) characterises linking error as the sum of the combined errors of item difficulty estimates obtained from two independently calibrated cycles Y1 and Y2 in which the trend items occur (potentially together with a set of items unique to each cycle). In other words, the linking error quantifies the item-by-cycle interactions, not the item-by-country-by-cycle interactions. The rationale for this approach was that the Rasch model is “symmetric,” which means an increase in difficulty of items can be compensated by the same increase in average ability.

This approach to estimating linking error assumes that the variability of item parameters over cycles directly translates into variability of person estimates, and that the average effect of parameter differences is zero, since the scales between Y1 and Y2 are linked. Thus, all country measures are affected in the same way by linking errors, which results in scale-level linking error. Moreover, note that there are two sets of trend item parameter estimates for each cycle, but neither is correct because both differ from the expected true parameters.

Other contributing factors to linking error are limited sample sizes and the number of unique items in each assessment cycle (unique means only administered in a particular cycle). In turn, this variability stems from differences in the calibration sample and the sampling variability associated with choosing a calibration sample, and from the presence of items that are unique to each cycle. This uncertainty is also related to the particular sample of trend items that were used in both cycles.

The above approach is only possible for the Rasch model, as there is only one parameter type incorporated in the linking error. In addition, it does not directly take into account the differences due to model error, for example, differential item functioning across countries that is not fully accounted for in modelling. Therefore, a new approach to characterise linking error was implemented in PISA 2015 that provides an estimate of the expected uncertainty due to differences between older and newer calibrations with more data.

The premise underlying the new approach is consistent with previous PISA cycles, yet it makes a different set of assumptions that can also be applied to more general IRT model-based linking. As in past cycles, scale-level differences across countries for adjacent calibrations are considered as the target of inference. The effect of the variability of two calibrations is evaluated at the cross-country level, while within-country sampling variability is not targeted. Moreover, sampling variance and measurement error are two separate variance components that are accounted for by plausible values and replicate weights-based variance estimation. Taken together, the focus lies on the expected variability on the



country level over calibrations, which is the highest reporting level. The calibration differences incorporate scaling differences, model differences, and different sets of unique items that may lead to somewhat different estimates in the two calibrations that can be compared with regard to linking error.

The definition of calibration differences starts from the ability estimates of a respondent  $v$  from country  $g$  in a target cycle under two separate calibrations (e.g. the original calibration of a particular PISA cycle and its recalibration), C1 and C2. We can write for calibration C1:

**9.26**

$$\tilde{\theta}_{v,C1,g} = \theta_{v,true} + \hat{u}_{C1,g} + \tilde{e}_v$$

where  $\hat{u}_{C1,g}$  denotes the estimated country specific error term in C1 and  $\tilde{e}_v$  is the respondent specific measurement error; and for calibration C2 accordingly:

**9.27**

$$\tilde{\theta}_{v,C2,g} = \theta_{v,true} + \hat{u}_{C2,g} + \tilde{e}_v$$

Defined in this way, there may be country level differences in the expected values of respondents based on the calibration. These are a source of uncertainty and can be viewed as adding variance to country level estimates. Given the assumption of a country level variability of estimates due to C1 and C2 calibrations, for the differences between estimates we find:

**9.28**

$$\tilde{\theta}_{v,C1,g} - \tilde{\theta}_{v,C2,g} = \hat{u}_{C1,g} + \hat{u}_{C2,g}$$

And the expectation can be estimated by:

**9.29**

$$E(\hat{u}_{C1,g} - \hat{u}_{C2,g}) = \tilde{\mu}_{g,C1} - \tilde{\mu}_{g,C2} = \hat{\Delta}_{C1,C2,g}$$

Across countries, the expected differences of country means ( $\tilde{\mu}$ ) can be assumed to vanish since the scales are transformed after calibrations to match moments. That is, we may assume:

**9.30**

$$\sum_{g=1}^G E(\hat{u}_{C1,g} - \hat{u}_{C2,g}) = 0 = \sum_{g=1}^G \hat{\Delta}_{C1,C2,g}$$

The variance of the differences of country means based on C1 and C2 calibrations can then be considered the linking error of the trend comparing the Y2 cycle means that were used to obtain calibration C2 estimates, and the Y1 cycle estimates. The link error can be written as:

**9.31**

$$V[\hat{\Delta}_{C1,C2,g}] = \frac{1}{G} \sum_{g=1}^G (\tilde{\mu}_{g,C1} - \tilde{\mu}_{g,C2})^2$$

The main characteristics of the new approach can be summarised as follows:

- Scale-level differences across countries from adjacent-cycle IRT calibrations C1 and C2 are considered.
- The effect of the variability of scale-level statistics between two calibrations is evaluated at the country level.
- Within-country sampling variability is not targeted.
- Sampling variance and measurement error are two separate variance components that are accounted for by plausible values and replicate weights-based variance estimation.



The use of this variance component is analogous to that of previous cycle linking errors. The variance calculated in (9.31) is a measure of uncertainty due to re-estimation of the model when using additional data from subsequent cycles, obtained with potentially different assessment designs, estimation methods, and underlying databases. In the application to linking error estimation for the 2015 PISA trend comparisons, a robust version of the scale was used. The robust measure of standard deviation that was used is the  $S_n$  statistic (Rousseeuw and Croux, 1993). It is defined as:

**9.32**

$$S_n = 1.1926 * \text{med}_i \left( \text{med}_j (|x_i - x_j|) \right)$$

The differences defined above are plugged into the formula, that is,  $x_i = \hat{\Delta}_{C1,C2,i}$  are used to calculate the linking error for comparisons of cycles Y1 and Y2 based on calibrations C1 (using only Y1 data) and C2 (using Y2 data and additional data including Y1). The robust estimates of linking error between cycles, by domain are presented in Chapter 12.

The  $S_n$  statistic is available in SAS as well as the R package “robustbase”. See also <https://cran.r-project.org/web/packages/robustbase/robustbase.pdf>. The  $S_n$  statistic was proposed by Rousseeuw and Croux (1993) as a more efficient alternative to the scaled median absolute deviation from the median (1.4826\*MAD) that is commonly used as a robust estimator of standard deviation.

## Population modelling in PISA 2015

The population model described earlier (*Latent regression model and population modelling*) was applied to the PISA 2015 data after the IRT scaling in order to generate 10 plausible values for each student. Plausible values for students reflect the information contained in responses to the items of domains that respondents actually took and the context questionnaire variables. Plausible values in all the major domains were generated for all students participating in the assessment, regardless of whether they were administered items in that domain. In addition, in countries where collaborative problem solving was administered plausible values were generated for all students, regardless of the test form they took. That is, respondents will be assigned plausible values for domains in which they did not participate, borrowing statistical information from students similar in performance on other domains, and in their responses to background data. This is enabled through the use of the population model, which uses the covariance information among all domains and also nearly all context questionnaire variables, as well as data about the number of not-reached items and other variables relevant to predicting proficiency distributions within each country.

Students who received plausible values for the domain(s) they did not take, but these values have a larger uncertainty (measurement error) than the plausible values for the other domains (that were administered to them). The measurement error has to be taken into account when dealing with the plausible values in secondary analyses. By using repeated analysis with each of the 10 plausible values, the measurement error will already be reflected in the analyses, and the final aggregation of results can be conducted in a way that the variability across the 10 analyses is properly reflected.

The following sections provide information about how the population model was applied to PISA 2015 data, how plausible values were generated, and how plausible values can be used in further analyses.

### Treatment of students with fewer than six test item responses

This section addresses the issue of students who provided background information but did not completely respond to the test items. A minimum of six completed items per domain was necessary to assure sufficient information about the proficiency of students. In general, there are very few students<sup>3</sup> (0.04%) with responses to fewer than six test items in at least one of the core test domains (reading, mathematics, science and collaborative problem solving). These cases, identified across the core domains, were initially removed from the first round of the population modelling for the core domains as well as for financial literacy. More precisely, students with responses to fewer than six test items per domain were not included in a first run of the population modelling (with regard to the regression model) in order to obtain unbiased  $\Gamma$  and  $\Sigma$ . In a second analysis step of population modelling, the regression parameters were treated as fixed to obtain plausible values for all cases, including those with fewer than six responses to test items.

For the science domain, students had to respond to at least six items in one of the subscales within a science dimension or subscale group (competency, system, knowledge) to be included in the latent regression model (note that a population modelling was done on the level of scientific subscale dimensions).

In PISA 2015 all consecutively missing responses at the end of a cluster were treated as “not reached” and thus coded as missing response (similar to “not administered” items); hence, they were ignored in the model. This scoring method is important with regard to the population model described (*Data yield and data quality*) since the population model is based on responses to the background questions and the test items.

### **Handling of item-by-country/language interactions**

The population model was estimated separately for each country, with the exceptions of Belgium (Dutch, French), Canada (English, French), Israel (Hebrew, Arabic), and Qatar (Arabic, English) where the model was estimated separately by language. Item parameter files for test items, including common and unique item parameters, were obtained from the IRT scaling (described earlier in this chapter). Because the IRT scaling used a multiple-group (concurrent) calibration method, an item parameter file was created for each country. If there were larger language groups that allowed separate evaluation of item fit, these item parameter files were merged so that one file resulted for each country, except for Belgium, Canada, Israel, and Qatar, which received two separate item parameter files each (one for each main language); the language groups of those countries were introduced separately in the population modelling. By incorporating country-by-language group item parameter files into the analyses, the population modelling accounted for unique item parameters and thus for item-by-country and item-by language interactions.

The country-specific conditioning model assures that the latent regression is based only on data obtained within the same country version for background questionnaire and test (country-by-language where feasible). This ensures that the unique relationship between background variables and proficiency variables can be represented for each country without bias. The use of country-specific item parameter files that contain a large number of common/international item parameters ensures the comparability of the plausible values.

### **Population model for the domains**

To generate plausible values for the domains of reading, mathematics, science, financial literacy and collaborative problem-solving, multidimensional population models were used. The multidimensional models included reading, mathematics and science, collaborative problem solving (computer-based assessment mode only) and financial literacy (if available).

The plausible value variables for the domains follow the naming convention PV1xxxx through PV10xxxx, where “xxxx” takes on the following form:

- READ for reading
- MATH for mathematics
- SCIE for science
- CLPS for collaborative problem solving
- FLIT for financial literacy

### **Population model for the science subscales**

There were several subscales reported for Science. These were knowledge scales (content; and procedural and epistemic), competency subscales (explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically) and system subscales (physical; living; and earth and space).

To generate plausible values for the science subscales, multidimensional population models were used. In total, three different multidimensional population models were used within each country:

- model 1: reading, mathematics, collaborative problem solving (computer based assessment mode only) and the science knowledge subscales
- model 2: reading, mathematics, collaborative problem solving (computer-based assessment mode only) and the science competency subscales
- model 3: reading, mathematics, collaborative problem solving (computer-based assessment mode only) and the science system subscale.

The aim of generating plausible values for the different science subscales, is to represent a more nuanced picture of important aspects within the overall science framework. These subscales allow for investigations of different aspects within the science domain, thus, exploring further the variability of skills across participating countries. Table 9.25 gives



an overview of the distributions of 85 trend and 99 new items (184 in total) to the three scales knowledge, competency, and system as well as the eight subscales: content; procedural and epistemic; explain phenomena scientifically; evaluate and design scientific inquiry; interpret data and evidence scientifically; physical; living; earth and space. It should be noted that the three science subscales types are based on a three-way classification of the same 184 items (distributed into the 2+3+3=8 subscales) and thus cannot be compared among each other, since these contain the same items, classified in three different ways.

**Table 9.25 Distribution of 85 trend and 99 new items to the science scales and subscales**

Knowledge			Competency			System		
Subscales	Trend	New	Subscales	Trend	New	Subscales	Trend	New
Content	51	47	Explain phenomena scientifically	41	47	Physical	28	33
Procedural and epistemic (merged)	34 (24+10)	52 (36+16)	Evaluate and design scientific enquiry	16	23	Living	39	35
			Interpret data and evidence scientifically	28	29	Earth and space	18	31
Total no. of trend/new items	85	99	Total no. of trend/new items	85	99	Total no. of trend/new items	85	99
Total no. of items	184		Total no. of items	184		Total no. of items	184	

Note: After the population modelling was finished and results reported to countries, the science experts recommended the reclassification of one item from the subscale "interpret data and evidence scientifically" to the subscale "explain phenomena scientifically" (see Chapter 2 for an updated item table). This change will be addressed in future PISA cycles but is not reflected in the PISA 2015 analyses.

The information about the eight subscales (2+3+3 subscales) was included in the population modelling. For example, the population model for scientific knowledge included the information about which items belonged to the two subscales "content" and "procedural and epistemic." Please note that for science, three additional population models (one for each of the three classifications of items) were computed in addition to science as a main scale. However, 10 plausible values were generated for each of the eight subscales.

The plausible value variables for the Science subscales follow the naming convention PV1xxxx through PV10xxxx, where "xxxx" takes on the following form:

- SKCO Science subscale – Content (knowledge)
- SKPE Science subscale – Procedural and epistemic (knowledge)
- SCEP Science subscale – Explain phenomena scientifically (competency)
- SCED Science subscale – Evaluate and design scientific inquiry (competency)
- SCID Science subscale – Interpret data and evidence scientifically (competency)
- SSPH Science subscale – Physical (system)
- SSLI Science subscale – Living (system)
- SSES Science subscale – Earth and science (system)

### Generating plausible values

Plausible values are multiple imputations of proficiency values based on information from the test items and information provided by the students in the background context questionnaire (BQ). Plausible values are used to obtain more accurate estimates of group proficiency than would be obtained through an aggregation of point estimates. A more detailed description is given in *Latent regression model and population modelling* above as well as in Mislevy (1991), Thomas (2002), and von Davier, Sinharay, Oranje, and Beaton (2006).

In PISA, the computation of group-level reporting statistics – involving scores in each of the domains (reading, mathematics, science, financial literacy and collaborative problem solving) as well as science subscales – is based on 10 independently drawn plausible values for each of the test domains and subscales for each student. Each set of plausible values is equally well designed to estimate population parameters; however, multiple plausible values are 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 earlier in this chapter (in *Latent regression model and population modelling* and *Analysis of data with plausible values*).

For the population modelling and the calculation of plausible values for the scales of PISA, the computer programme DGROUP (Rogers et al., 2006) was used.

A normal multivariate distribution was assumed for  $P(\theta_j | x_j, y_j, \Gamma, \Sigma)$ , with a common variance,  $\Sigma$ , and with a mean given by a linear model with slope parameters,  $\Gamma$ , based on the principal components of several hundred selected main effects from the vector of context questionnaire variables.

The item parameters for the test items were obtained from the concurrent item calibration described earlier in this chapter (see *The IRT models for scaling, National and international item calibration and Handling of item-by-country/language and item-by-mode interactions*) using the data from past PISA cycles (2006-2012) and PISA 2015 as described above. The result of the concurrent calibration is a scale that provides comparable results across the different PISA cycles. To calculate the plausible values for PISA 2015 only, the item parameters for items administered in PISA 2015 were used in the population modelling.

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 of this report. All variables in the context questionnaire were contrast coded before they were processed further in the population model. Contrast coding allows for the inclusion of codes for refused responses, avoiding the necessity of linear coding. The contrast coding scheme is reproduced in Annex B. 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. 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 cognitive skills. In PISA, each set of principal components  $y^c$  (or conditioning variables) was selected to include 80 percent of the variance, or not to exceed a number of principal components greater than 5% of the raw sample size, with the aim of explaining as much variance as possible while at the same time avoiding over parameterization of the model.

Principal component scores based on nearly all (contrast coded) background variables were used in PISA, including international variables (collected by every participating country) as well as national background variables (country-specific variables in addition to the international variables).

Students with responses to fewer than six test items per domain were not included in a first run of the regression model in order to obtain unbiased  $\Gamma$  and  $\Sigma$ . In a second analysis step of population modelling, the regression parameters were treated as fixed to obtain plausible values for all cases, including those with fewer than six responses to items (see earlier section *Treatment of students with fewer than six cognitive item responses* for more information).

The financial literacy plausible values for the students who took this domain are based on a latent regression model that included the general background questionnaire variables plus the additional financial literacy background questions that were administered together with the financial literacy test items. A separate latent regression model based on the general background questionnaire variables alone was used for the remaining students who did not take the financial literacy test items as well as the financial literacy background questions.

Students received plausible values for each test domain administered in their country according to the test design that applied in a particular country (paper- versus computer-based assessment, financial literacy selected or not; collaborative problem solving selected or not; see Chapter 2 for more information on the test design). This means, students also received plausible values for test domains that were not administered to them. The same applies to students who took the Une Heure (UH) test design.



## Note

1. A subset of cases from certain countries had to be excluded from the IRT calibration due to adjudication and data quality issues (please see Chapter 14 for more information).
2. Note that the random effect in Model 9.16 could be adjusted for each country separately, so this model picks up country differences as well. The similarity between the fit of models 9.16 and 9.15 shows, that no country-specific constraints are needed.
3. Note that a student was only considered a “respondent” and given an analysis weight if he or she responded to at least one test item and a certain amount of the context questionnaire items, or if he or she responded to at least half of the test items in cases of providing no context questionnaire information.

## References

- 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.
- 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.
- Allen, N. L., J. R. Donoghue and T. L. Schoeps** (2001), *The NAEP 1998 Technical Report*, NCES 2001-509, Office of Educational Research and Improvement, National Center for Education Statistics, U.S. Department of Education, Washington, DC: National Center for Education Statistics.
- 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.
- Bock, R. D. and M. F. Zimowski** (1997), “Multiple group IRT”, In W. J. van der Linden and R. K. Hambleton (eds.), *Handbook of Modern Item Response Theory* (pp. 433-448), Springer-Verlag, New York, NY.
- Gilula, Z. and S. J. Haberman** (1994), “Conditional log-linear models for analyzing categorical panel data”, *Journal of the American Statistical Association*, Vol. 89/426, pp. 645-656.
- Glas, C. A. W. and K. Jehangir** (2014), “Modelling 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.
- Holzinger, K. and F. Swineford** (1937), “The bi-factor method”, *Psychometrika*, 2, 41-54.
- Johnson, E. G.** (1989), “Considerations and techniques for the analysis of NAEP data”, *Journal of Educational Statistics*, Vol. 14/4, pp. 303-334.
- Johnson, E. G., and K. F. Rust** (1992), “Population inferences and variance estimation for NAEP data”, *Journal of Educational Statistics*, Vol. 17, pp. 175-190.
- Kreiner, S. and K. B. Christensen** (2014), “Analyses of model fit and robustness: A new look at the PISA scaling model underlying ranking of countries according to reading literacy”, *Psychometrika*, Vol. 79/2, pp. 210-231.
- Leys, C. et al.** (2013), “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median”, *Journal of Experimental Social Psychology*, Vol. 49, pp. 764-766.
- Little, R. J. A. and D. B. Rubin** (1983), “On jointly estimating parameters and missing data”, *American Statistician*, Vol. 37, pp. 218-220.
- Martin, M. O., K. D. Gregory and S. E. Stemler**, (eds.) (2000), *TIMSS 1999 Technical Report*, International Study Center, Boston, MA.
- Masters, G. N.** (1982), “A Rasch model for partial credit scoring”, *Psychometrika*, Vol. 47, pp. 149-174.
- 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.
- Mazzeo, J. and M. von Davier** (2008), *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*, Doc. ref. EDU/PISA/GB(2008)28, Retrieved from <http://www.oecd.org/dataoecd/44/49/41731967.pdf>.
- Meredith, W** (1993), “Measurement invariance, factor analysis and factorial invariance”, *Psychometrika*, Vol. 58, pp. 525-543.

- Mislevy, R. J.** (1991), "Randomization-based inference about latent variables from complex samples", *Psychometrika*, Vol. 56/2, pp. 177-196.
- Mislevy, R. J.** (1985), "Estimation of latent group effects", *Journal of the American Statistical Association*, Vol. 80/392, pp. 993-997.
- Mislevy, R. J.** et al. (1992), "Estimating population characteristics from sparse matrix samples of item responses", *Journal of Educational Measurement*, Vol. 29, pp. 133-161.
- 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 **von Davier, M.** (2014), "Toward increasing fairness in score scale calibrations employed in international large-scale assessments", *International Journal of Testing*, Vol. 14/1, pp. 1-21, doi:10.1080/15305058.2013.825265.
- Oliveri, M. E.** and **von Davier, M.** (2011), "Investigation of model fit and score scale comparability in international assessments", *Psychological Test and Assessment Modelling*, Vol. 53/3, pp. 315-333, Retrieved from [http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011\\_20110927/04\\_Oliveri.pdf](http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf).
- Rasch, G.** (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen, Denmark: Nielsen and Lydiche (Expanded edition, Chicago, University of Chicago Press, 1980).
- Rogers, A.** et al. (2006), DGROUP (computer software), Educational Testing Service, Princeton, NJ.
- Rosenbaum, P. R.** (1988), "Permutation tests for matched pairs with adjustments for covariates", *Applied Statistics*, Vol. 37, pp. 401-411.
- Rousseeuw, P. J.** and **C. Croux** (1993), "Alternatives to the median absolute deviation", *Journal of the American Statistical Association*, Vol. 88/424, pp. 1273-1283, doi:10.2307/2291267, JSTOR 2291267.
- Rubin, D. B.** (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons, New York, NY.
- Rust, K. F.** (2014), "Sampling, weighting, and variance estimation 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*, pp. 117-154, CRC Press, Boca Raton, FL.
- Sireci, S. G., D. Thissen, and H. Wainer** (1991), "On the reliability of testlet-based tests", *Journal of Educational Measurement*, Vol. 28, pp. 237-247.
- Skrondal, A.** and **S. Rabe-Hesketh** (2004), *Generalized Latent Variable Modelling: Multilevel, Longitudinal and Structural Equation Models*, Chapman and Hall/CRC, Boca Raton, FL.
- Thomas, N.** (2002), "The role of secondary covariates when estimating latent trait population distributions", *Psychometrika*, Vol. 67/1, pp. 33-48.
- Thomas, N.** (1993), "Asymptotic corrections for multivariate posterior moments with factored likelihood functions", *Journal of Computational and Graphical Statistics*, Vol. 2, pp. 309-322.
- van der Linden, W. J.** and **R. K. Hambleton** (2016), *Handbook of Modern Item Response Theory*, 2<sup>nd</sup> ed. Springer, New York, NY.
- von Davier, M.** (2016), "The Rasch Model: Chapter 3", in van der Linden, W. (ed.) *Handbook of Item Response Theory*, Vol. 1, Second Edition, CRC Press, pp. 31-48.
- 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.** 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., 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](http://www.ierinstitute.org/IERI_Monograph_Volume_02_Chapter_01.pdf).
- von Davier, M. and A. von Davier** (2007), "A unified approach to IRT scale linking and scale transformations", *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, Vol. 3/3, pp. 115-124.
- von Davier, M. and K. Yamamoto**, (2007), "Chapter 6: Mixture distribution Rasch models and Hybrid Rasch models", in: M. von Davier and C.H. Carstensen, *Multivariate and Mixture Distribution Rasch Models*, Springer, New York.



**von Davier, M.** et al. (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, Psychometrics*, Vol. 26, 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 Measurement*, Vol. 28/6, pp. 389-406.

**Wainer, H., E. T. Bradlow** and **X. Wang** (2007), *Testlet response theory and its applications*. Cambridge University Press, New York, NY.

**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.

**Wilson, M.** and **R. J. Adams**, (1995), "Rasch models for item bundles", *Psychometrika*, Vol. 60, pp. 181-198.

**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.

**Wise, S. L.** and **C. E. DeMars** (2005), "Low student effort in low-stakes assessment: Problems and potential solutions", *Educational Assessment*, Vol. 10/1, pp. 1-17.

**Wu, M. L., R. J. Adams** and **M. R. Wilson**, (1997), ConQuest: Multi-Aspect Test Software [computer program]. Camberwell, Australia: Australian Council for Educational Research.

**Yamamoto, K.** (1998) "Scaling and scale linking", in T. S. Murray, I. S. Kirsch and L. B. Jenkins (eds.), *Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey* (pp. 161-178), National Center for Education Statistics, Washington, DC.

**Yamamoto, K.** (1997), 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.





10

# Data management procedures

<b>Introduction .....</b>	188
<b>Data management at the international and national level .....</b>	188
<b>The data management process and quality control .....</b>	190
<b>Harmonisation.....</b>	194
<b>Validation.....</b>	195
<b>Scoring and derivation .....</b>	195
<b>Deliverables .....</b>	196



## INTRODUCTION

In PISA, as in any international survey, a set of standard, data collection requirements guides the creation of an international database that allows for valid within-and-cross-country comparisons and inferences to be made. For both paper-based (PBA) and computer-based (CBA) assessments, these standard requirements are developed with three major goals in mind: consistency, precision and generalisability. In order to support these goals, data collection and management procedures are applied in a common and consistent way across all participants' data to ensure data quality. Even the smallest errors in data capture, coding, and/or processing may be difficult, if not impossible, to correct; thus, there is a critical need to avoid or at the very least minimise the potential for errors.

Although these international standards and requirements stipulate a collective agreement and mutual accountability among countries and contractors, PISA is an international study that includes countries with unique educational systems and cultural contexts. The PISA standards provide the opportunity for participants to adapt certain questions or procedures to suit local circumstances, or add components specific to a particular national context. To handle these national adaptations, a series of consultations was conducted with the national representatives of participating countries in order to reflect country expectations in agreement with PISA 2015 technical standards. During these consultations, the data coding of the national adaptations to the instruments was discussed to ensure their recoding in a common international format. The guidelines for these data management consultations and recoding concerning national adaptations are described later on in this chapter.

An important part of the data collection and management cycle is not only to control and adapt to the planned deviations from general standards and requirements, but also to control and account for the unplanned and/or unintended deviations that require further investigation by countries and contractors. These deviations may compromise data quality and/or render data corrupt, or unusable. For example, certain deviations from the standard testing procedures are particularly likely to affect test performance (e.g. session timing, the administration of test materials, and tools for support such as rulers and/or calculators). Sections of this chapter outline aspects of data management that are directed at controlling planned deviations, preventing errors, as well as identifying and correcting errors when they arise.

Given these complexities – the PISA timeline and the diversity of contexts in the administration of the assessment – it remains an imperative task to record and standardise data procedures, as much as possible, with respect to the national and international standards of data management. These procedures had to be generalised to suit the particular cognitive test instruments and background questionnaire instruments used in each participating country. As a result, a suite of products are provided to countries that include a comprehensive data management manual, training sessions, as well as a range of other materials, and in particular, the data management software designed to help National Project Managers (NPMs) and National Data Managers (NDMs) carry out in a consistent way data management tasks, prevent introduction of errors, and reduce the amount of effort and time in identifying and resolving data errors.

This chapter summarises these data management quality control processes and procedures and the collaborative efforts of contractors and countries to produce a final database for submission to the OECD.

## DATA MANAGEMENT AT THE INTERNATIONAL AND NATIONAL LEVEL

### Data management at the international level

To ensure compliance with the PISA technical standards, the following procedures were implemented to ensure data quality in PISA 2015:

- standards, guidelines, and recommendations for data management within countries
- data management software, manuals and codebooks for National Centres
- hands-on data management training and support for countries during the national database building
- management, processing, and cleaning for data quality and verification at the international and national level
- preparation of analysis and dissemination of databases and reports for use by the contractors, OECD and the National Centres
- preparation of data products (e.g. Data Explorer, IDB Analyser) for dissemination to contractors, National Centres, the OECD, and the scientific community.



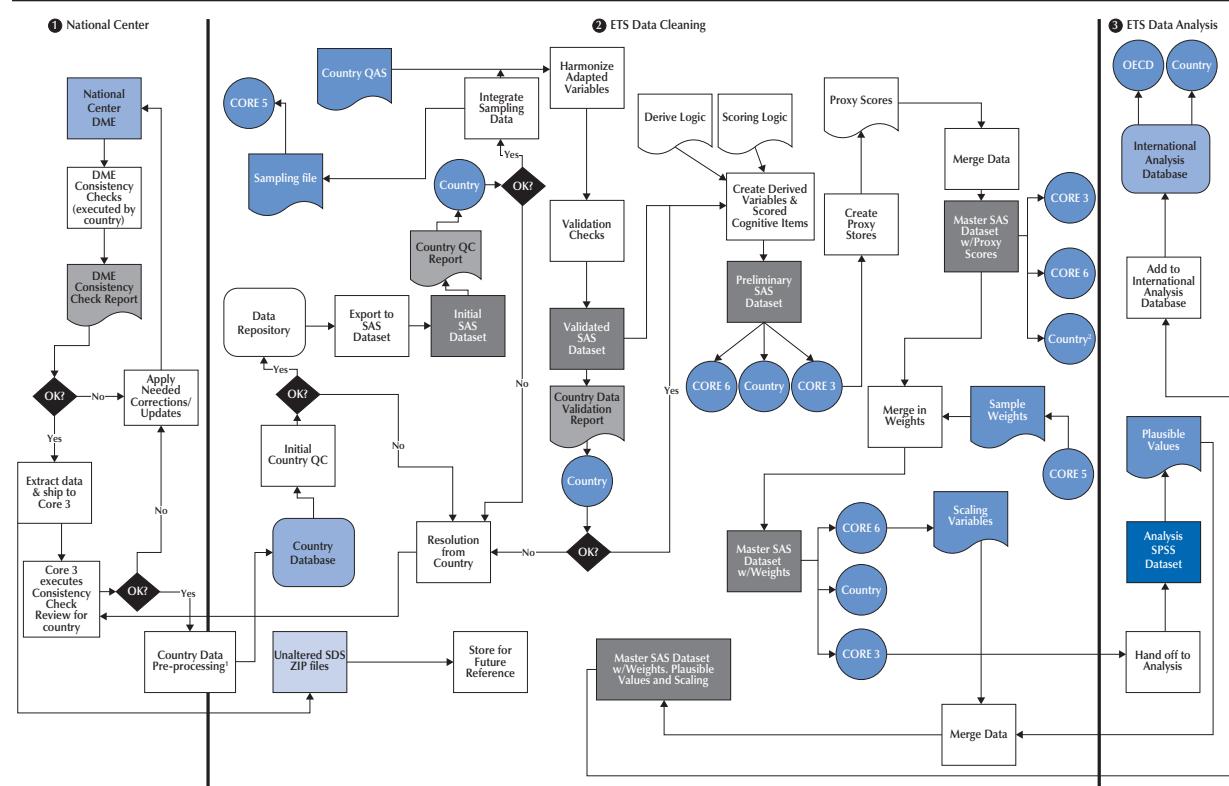
ETS Data Management and Analysis had overall responsibility for data management and relied on the following for information and consultation:

- ETS Project Management (Core 2 and Core 7): ETS Project Management provided contractors with overview information on country specifics including national options, timelines and testing dates, and support with country correspondence and deliverables planning.
- DIPF (Core 6): As the Background Questionnaire (BQ) experts, DIPF provided BQ scaling and indices, BQ data, support for questionnaire workflows and negotiations with National Centres concerning questionnaire national adaptations, harmonisation review, and BQ derived variables.
- Westat (Sampling) (Core 5): Leading the sampling tasks for PISA, Westat provided review and quality control support with respect to sampling and weighting. Westat is instrumental in providing guidance for quality assurance checks with regard to national samples.
- Westat (Survey Operations) (Core 4): Key to the implementation of the PISA assessment in countries, Westat's Survey Operations team supported countries through the PISA 2015 cycle. In addition to organising PISA meetings, Westat was responsible for specific quality assurance of the implementation of the assessment and submission of data to the National Centres.
- OECD: The OECD provided support and guidance to all contractors with respect to the specific area of expertise. The OECD's review of data files and preliminary data products provided the ETS Data Management and Analysis teams with valuable information in the structure of the final deliverables.

## Data management at the national level

As the standards for data collection and submission involve a series of technical requirements and guidelines, each participating country appointed a National Project Manager, or NPM, to organise the survey data collection and management at the National Centre. NPMs are responsible for ensuring that all required tasks, especially those relating

■ Figure 10.1 ■  
Overview of the data management process



1. Additional checks on data; data recovery processing; rescore of cognitive items; timing data extraction; coding reliability export for analysis.  
2. Interim databases delivery to country included proxy scopes.

to the production of a quality national database, are carried out on schedule and in accordance with the specified international standards and quality targets. The NPM is responsible for supervising, organising and delegating the required data management<sup>1</sup> tasks at the national level. Further, as these data management tasks require more technical skills of data analysis, NPMs were strongly recommended to appoint a National Data Manager (NDM) to complete all tasks on time and supervise support teams during data collection and data entry. These technical tasks for the NDM included, but were not limited to: collaborating with ETS on template codebook adaptations; integration of data from the national PISA data systems; manual capture of data after scoring; export/import of data required for coding (e.g. occupational coding); and data verification and validation with a series of consistency and validity checks. In order to adhere to quality control standards, one of the most important tasks for National Centres concerned data entry and the execution of consistency checks from the primary data management software, the PISA Data Management Expert, or "DME." For PISA 2015, Figure 10.1 provides the workflow of the data management process.

The next section outlines the data management process as well as the application of additional quality assurance measures to ensure proper handling and generation of data. Additionally, more information is provided on the PISA 2015 DME as well as the phases of the data management cleaning and verification process.

## THE DATA MANAGEMENT PROCESS AND QUALITY CONTROL

The collection of student, teacher and test administrator responses on a computer platform into electronic data files provided an opportunity for the accurate transcription of those responses and the collection of process data, including response actions and timing. It also presented a challenge to develop a system that accepted and processed these files and their variety of formats as well as supporting the manual entry of data from paper forms and booklets. To that end, the Data Management team acquired a license for the adaptation, use, and support of the Data Management Expert (DME) software, which had previously proved successful in the collection and management of the data for the survey for adult skills (PIAAC) under a separate contract.

The DME software is a high performance .NET based, self-contained application that can be installed on most Windows operating systems (Windows XP or later), including Surface Pro and Mac Windows, and does not require an internet connection to operate. It operates on a separate database file that has been constructed according to strict structural and relational specifications that define the data codebook. This codebook is a complete catalogue of all of the data variables to be collected and managed and the arrangement of these variables into well-defined datasets that correspond to the various instruments involved in the administration of the assessment. The DME software validates the structure of the codebook part of the database file and, if successful, creates the data tables within the same file for the collection and management of the response and derivative data.

With this process, the Data Management contractor first developed and tested a template of the international data codebook representing all the data to be collected across CBA and PBA countries without national adaptations. The datasets in this codebook also included those for all international options (such as financial literacy, teacher questionnaires, etc.) regardless of each country's mode or selected options. The national templates for each of the CBA countries are built upon this international template, using the questionnaire adaptations coded in the Questionnaire Adaptation Tool (QAT) and removing the datasets for PBA countries and the international options not implemented in the country. The national templates for each of the PBA countries consist of the international template with the CBA-specific datasets removed. The National Data Manager (NDM) for each PBA country is trained on and is responsible for implementing and testing the national adaptations to the delivered codebook.

The DME software provided three modes of entering data into the project database: imports of standard format files, imports of PISA specific archive files, and direct manual entry from paper forms and booklets. The standard format files are either Excel workbooks or CSV files and include such data as the results of the occupational coding. The PISA-specific files include the archive files that are generated by the student delivery system (SDS) software at the student level and the school and teacher questionnaire data files that are downloaded from the questionnaire web site by each NDM. The identification and extraction of data from these sources requires special programming within the DME software and supporting tables within the codebook files.

PBA countries performed direct manual entry into the system from paper forms and booklets. PBA data managers were required to program the codebook with the appropriate variables based on the booklet number and according to data management guidelines. Data entry was also required for the Parent Questionnaire when that option was selected by both PBA and CBA countries. An important feature of the DME software is the ability to create multiple copies of



the project codebook for use on remote computers and to merge the databases created on each remote site into the master project database. This permits the establishment of a manageable processing environment based on a common codebook structure to guarantee the accurate and consistent transcription of the data.

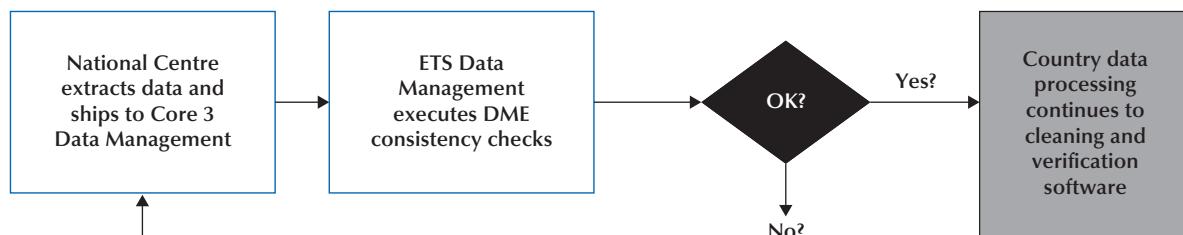
The DME software can also produce a series of reports at any point during data collection, including: detection of records with the same identification information, validation of all data values against the codebook specifications, and a set of consistency checks defined and coded by the Data Management contractor. These checks provided information on the completeness of the data across datasets, identified inconsistent responses within each questionnaire, and reported on the overall status of the data collection process. At the conclusion of data collection and processing in each country, the NDM was required to either resolve or explain the discrepancies uncovered by these reports and submit the annotated reports along with the final database to the Data Management contractor.

### **Pre-processing**

When data were submitted to the Data Management contractor, a series of pre-processing steps were performed on the data to ensure completeness of the database and accuracy of the data. Running the DME software was one of the first consistency checks on the data submission. In the field, National Centres were required to run these checks frequently for data quality and consistency. Although National Centres were required to execute these checks on their data, the Data Management contractor also executed these DME consistency checks in early data processing as a quick and efficient way to verify data quality.

These checks in addition of other internal checks for coding were executed and any inconsistencies were compiled into a report and returned to the National Centre for more information and/or further corrections to the data. If necessary, National Centres resubmitted their data to the Data Management contractor for any missing or incorrect information and document any changes made to the database in the consistency check report file. When countries redelivered data, Data Management refreshed the existing database with the newly-received data from the National Centre and continued with the same pre-processing steps again – executing another series of consistency checks to be sure all necessary issues are resolved and/or documented. In this initial step of processing, returning data inconsistencies to the National Centres was an iterative process with some times up to 4-5 iterations of data changes/updates from the country. Once resolved, the data continued to the next phase of the internal process – loading the database into the cleaning and verification software.

■ Figure 10.2 ■  
**Overview of the delivery and pre-processing phase**



### **Initial database load into SQL server and the cleaning and verification software**

With the pre-processing checks complete, the country's database advanced to the next phase of the process – data cleaning and verification. To reach the high quality requirements of PISA technical standards, the Data Management contractor created and used a processing software that merged datasets in SAS, but had the ability to produce both SAS and SPSS datasets. During processing, one to two analysts independently cleaned country databases, focusing on one country at a time in order to complete all necessary phases of quality assurance, in order to produce both SAS and SPSS datasets to the country and other contractors.

The first step in this process was to load the DME database onto the ETS Data Management cleaning and verification server. With the initial load of the database, specific quality assurance checks were applied to the data. These checks ensured:

- The project database delivered by the country used the most up-to-date template provided by the Data Management team which included all necessary patch files applied to the database. For PISA 2015, patch files were released by ETS

Data Management and applied to the SQL database by the National Data Manager to correct errors in the codebook or to modify the consistency checks in the DME software. For example, a patch may be issued if an item was misclassified as having 4 category response options instead of 5.

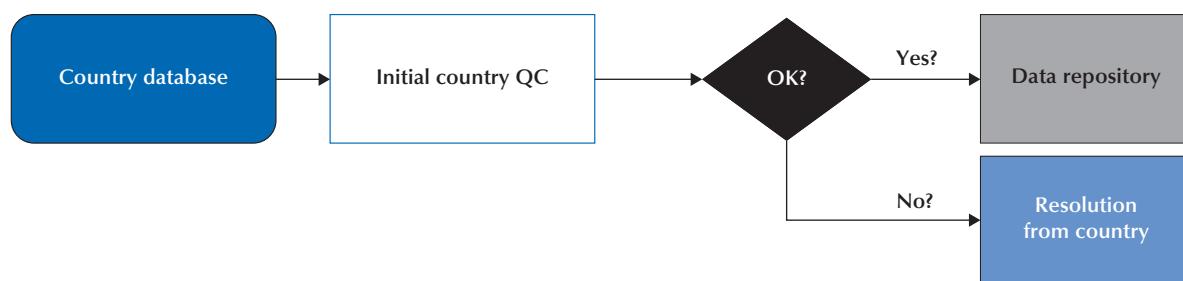
- The country database had the correct profile as dictated by the country's selected international options (e.g. Financial Literacy, UH booklet, etc.).
- The number of cases in the data files by country/language were in agreement with the sampling information collected by Westat.
- All values for variables that used a value scheme were contained by that value scheme. For example, a variable may have the valid values of 1, 3 and 5; yet, this quality assurance check would capture if an invalid value, e.g. "4", was entered in the data.
- Valid values that may have been miskeyed as missing values were verified by the country. For example, valid values for a variable might range from "1" to "100" and data entry personnel may have mistakenly entered a value of "99", intending to issue a value of "999". This is common with paper-based instruments. Each suspicious data point was investigated and resolved by the country.
- Response data that appeared to have no logical connection to other response data (e.g. school/parent records possessing no relation to any student records) were validated to ensure correct IDs are captured.

## Integration

After the initial load into the data repository and completion of early processing checks (Figure 10.3), the database entered the next phase of processing: Integration (Figure 10.4). During this integration phase, data which was structured within the country project database to assist in data collection was restructured to facilitate data cleaning. At the end of this step, a single dataset was produced representing each of the respondent types: student, school, and teacher (where applicable). Additionally, parent questionnaire data was merged with their child/student data.

■ Figure 10.3 ■

### Initial load of the National Centre database into SQL server for processing



In the main survey, the integration phase was a critical juncture because data management was able to analyse the data collected within the context of the sampling information supplied by the sampling contractor, Westat. Using this sampling information –captured in the Student Tracking Form – extensive quality control checks were applied to the data in this phase. Over 80 quality assurance checks were performed on the database during this phase, including specific checks such as: verifying student data discrepancies of students who are marked as present but do not have test or questionnaire data; students who are not of the “allowable” PISA age; and students who are marked absent but have valid test or questionnaire data. As a result of these quality assurance checks, a quality control report was generated and delivered to countries to resolve outstanding issues and inconsistencies. This report was referred to as the Quality Control (“Country QC”) Report.

In this report, ETS Data Management provided specific information to countries, including the name of the check and the description of the check as well as specific information, such as student IDs, for the cases that proved to be inconsistent or incorrect against the check. These checks included (but were not limited to):

- FORMCODE was blank or not valid.
- Student was missing key data needed for sampling and processing.
- Student was not in the allowable age.



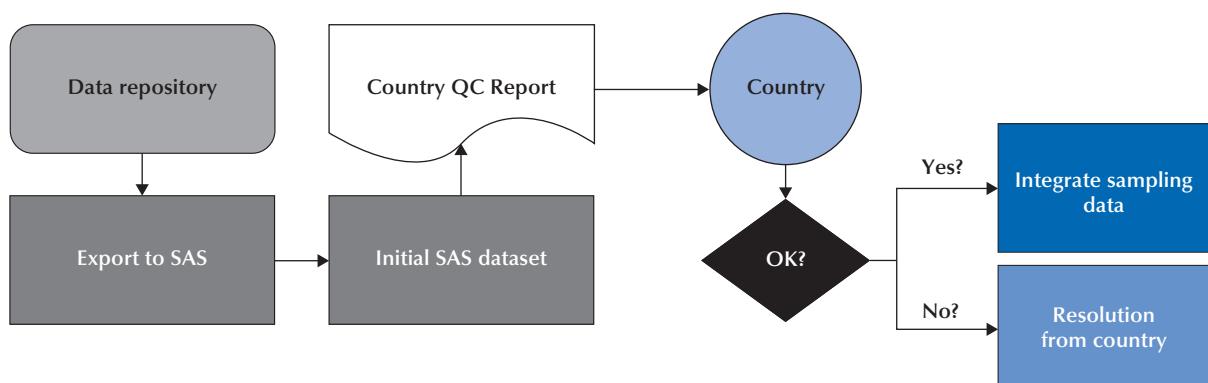
- Student was not represented in the STF.
- Students who were marked absent yet had records.
- Mother's or father's occupation appeared invalid or needed clarification because it was not of length 4.
- Student's grade was lower than expected.
- On the Teacher Questionnaire, a teacher was marked as a "non-participant"<sup>2</sup>, yet data existed.

In addition to quality control reporting, a series of important data processing steps occurred during integration.

- **Item Cluster Analysis:** For the purposes of data processing, it is often convenient to disaggregate a single variable into a collection of variables. To this end, a respondent's single booklet number was interpreted as a collection of Boolean variables which signalled the item clusters that the participant was exposed to by design. Similarly, the individual item responses for a participant were interpreted and coded into a single variable which represented the item clusters that the participant appears to have been presented. An analysis was performed which detects any disconnect between the student delivery system and the sampling design. Any discrepancies discovered were resolved by contacting the appropriate contractors.
- **Raw Response Data Capture:** In the case of paper-based administration, individual student selections (e.g. A, B, C, D) to multiple choice items were always captured accurately. This was not necessarily true, however, in the case of computer-based administrations. While the student delivery system captures a student's response, it fails to capture data in a format that could be used to conduct distractor analysis. The web-elements that are saved during a computer administration were therefore processed and interpreted into variables comparable to the paper-based administration.
- **Timing:** The student delivery system captured timing data for each screen viewed by the respondent. During the integration step, these timing variables were summed appropriately to give timing for entire sections of the assessment.
- **SDS Post-processing:** Necessary changes in the student delivery system (SDS) were sometimes detected after the platform was already in use. For example, a test item that was scored by the SDS may have had an error in the interpretation of a correct response, which was corrected in the SDS post-processing. These and other issues were resolved by the SDS developers and new scored response data was processed, issued, and merged by the Data Management team.

Following the Integration phase of data processing, the Country QC reports were generated and distributed to the National Centres. National Project Managers were asked to review the report and to address any reported violations. National Centres corrected or verified inconsistencies in the database from this report and returned the revised database to the Data Management contractor within a specific timeframe. Additionally, all data revisions were documented directly in the Country QC report for delivery to Data Management. After receiving the revised database, the Data Management team repeated the pre-processing phase to ensure no new errors were reported and, if no errors were found, the Data Management team re-executed the integration step. As with the pre-processing consistency checks phase, the integration step required several iterations and updates to country data if issues persisted and were not addressed by the National Centre. Frequently, one-on-one consultations were needed between the National Centre and the Data Management team in order to resolve issues. After all checks were revised and documented by the National Centre and no critical data violations remained, the data moved to the next phase in processing – i.e. national adaptation harmonisation.

■ Figure 10.4 ■  
**Integration process overview**



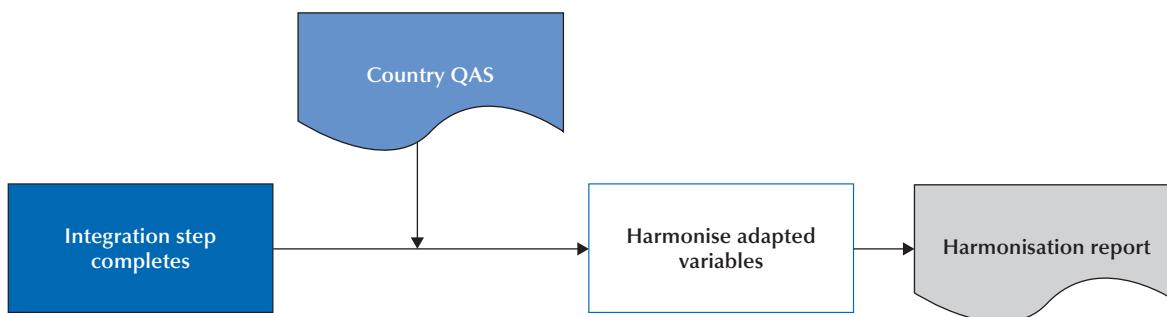
## HARMONISATION

### Overview of the workflow

As mentioned earlier in this chapter, although standardisation across countries was needed, countries had the opportunity to modify, or adapt, background questionnaire variable stems and response categories to reflect national specificities, referred to as “national adaptations”. As a result, changes to variables proposed by a National Centre occurred during the translation and adaptation process. National adaptations for questionnaire variables were agreed upon by the Background Questionnaire contractors. These discussions regarding adaptations happened in the “negotiation” phase between the country and the contractor as well as the translation verification contractor. All changes and adaptations to questionnaire variables were captured in the questionnaire adaptation sheet (QAS). It was the role of the Background Questionnaire contractor to use the country’s QAS file to approve national adaptations as well as any national adaptation requiring harmonisation code. The Data Management contractor also assisted the Background Questionnaire contractor in developing the harmonisation code for use in the cleaning and verification software. Throughout this process, it was the responsibility of the BQ contractor, with the assistance of the translation verification contractor, to ensure the QAS was complete and reflected the country’s intent and interpretation. Once adaptations were approved by the BQ contractor, countries were able to implement their approved national adaptations (using their QAS as a reference tool) in their questionnaire material. National Centres were required to document and implement all adaptations in the following resources: QAS and the DME.

Any issues surrounding the national adaptations were handled by the country as well as by both the BQ contractor and the Data Management contractor. Official BQ contractor approval of the harmonisation SAS code was required for data processing. Additionally, the BQ contractor was responsible for reviewing the harmonisation reports produced by ETS Data Management for any issues or concerns with national adaptations. The National Centres also reviewed these harmonisation reports and contacted both the BQ contractor and the Data Management contractor with any issues or changes. Changes were documented in the country QAS file. Following any change or modification, the data management team repeated the harmonisation stage in order to check the proposed changes.

■ Figure 10.5 ■  
Harmonisation process overview



### Harmonisation, or harmonised variables

In general, harmonisation or harmonising variables is a process of mapping the national response categories of a particular variable into the international response categories so they can be compared and analysed across countries. Not every nationally-adapted variable required harmonisation, but for those that required harmonisation, the Data Management team assisted the Background Questionnaire contractor with creating the harmonisation mappings for each country with SAS code. This code was implemented into the data management cleaning and verification software in order to handle these harmonised variables during processing.

Additionally, harmonisation consisted of adaptations for national variables where there was a structural change, e.g. question stem and/or variable response category options differ from the international version (this could be in the form of an addition or deletion of a response option and/or modification to the intent of the question stem or response option – as observed in variable SC013Q01TA where the country may alter the stem in creating a national adaptation and request information on the “type” of school in addition to whether the school is public or private). For example, more response categories may have been added or deleted; or perhaps two questions were merged (e.g. a variable may have five response options/choices to the question, but with the national adaptation the variable may have been modified to only have four response options/choices as only 4 make sense for the country’s purposes).



## VALIDATION

After the harmonisation process, the next phase in data cleaning and verification involved executing a series of validation checks on the data for contractor and country review.

### Validation overview

In addition to nationally-adapted variables, ETS Data Management collaborated with the BQ contractor to develop a series of validation checks that were performed on the data following harmonisation. Validation checks are consistency checks that provide National Centres with more detail concerning extreme and/or inconsistent values in their data. These violations of the validation checks were displayed in a validation report, which was shared with countries and contractors to observe these inconsistencies and make improvements for the next cycle of PISA. In the PISA 2015 main survey, National Centres did not make changes to revise these extreme and/or inconsistent values in the report. Rather, National Centres were instructed to leave the data as it is and make recommendations for addressing these issues in the data collection process during the next cycle of PISA. Although data modifications were not made for many of these validation checks, ETS Data Management required National Centres to document and provide more information into the nature of these data inconsistencies. Generally, validation checks of this nature captured inconsistent student, school and teacher data. For example, these checks may capture an inconsistency between the total number of years teaching to the number of years teaching at a particular school (TE0001); or an inconsistency in student data related to the number of class periods per week in maths and the allowable total class periods per week (ST059Q02TA). Throughout this PISA cycle, these validation checks often served as valuable feedback for data quality.

### Treatment of inconsistent and extreme values in PISA 2015 main survey data

During the preparations for the main survey international database release, some National Centres raised the issue of how to handle some extreme and/or inconsistent values within the data. The Data Management contractor, the Background Questionnaire contractor and the OECD agreed on implementing a specific approach to manage the extreme and/or inconsistent values present within the data.

Concerning the special handling of these inconsistent and/or extreme values, the following principles were followed:

- Support the results of DME software consistency checks from the PISA 2015 main survey. In most cases where there was an inconsistency, the question considered ‘more difficult’ was invalidated since this was more likely to have been answered inaccurately (for example, a question that involved memory recall or cognitive evaluation by the respondent).<sup>3</sup>
- Support the results of the validation checks from PISA 2015 main survey. In particular, it is key to note that cases that corresponded to selections from drop-down menus were not invalidated (for example, the variable, EC029Q01NA, from the Educational Career Questionnaire item, “How many years altogether have you attended additional instruction?”, however implausible).
- Apply stringent consistency and validity checks while computing derived variables.<sup>4</sup>

The specific range restriction rules for PISA 2015 are located in Figure 10.6 at the end of this chapter.

## SCORING AND DERIVATION

After validation, the next phase of data management processing involved parallel processes that occur with test data and questionnaire data:

- Scoring of test responses captured in paper booklets.
- Derivation of new variables from questionnaires.

### Scoring overview

The goal of the PISA assessment is to ensure comparability of results across countries. As a result, scoring for the tests was a critical component of the data management processing. While scores were generated for computer-based responses automatically, no such scoring variables existed for paper-based components. This step in the process was dedicated to creating these variables and inserting the relevant student responses. To aid in this process, the Data Management team implemented rules from coding guides developed by the Test Development team. The coding guides were organised in sections, or clusters, that outlined the value, or score, for responses. The Data Management team was not only responsible for generating the SAS code to implement these values, but was also responsible for implementing a series

of quality assurance checks on the data to determine any violations in scoring and/or any missing information. When missing scores were present in the data, the Data Management team consulted with the National Centre regarding these missing data. If National Centres were able to resolve these issues (e.g. student response information was mistakenly miscoded or not entered into the DME software), information was provided to the Data Management team through the submission of an updated, or revised, DME database and the necessary steps for pre-processing were completed. If the reported data inconsistencies were resolved, the scoring process was complete and the data proceeded to the next phase of processing.

The scoring variables also served as a valuable quality control check. If any items appeared to function not as expected (too difficult or too easy), further investigation was carried out to determine if a booklet printing error occurred or if systematic errors were introduced during data entry.

### **Derived variables overview**

The SAS derived variable code was generated by the BQ contractor, DIPF, for implementation into the Data Management cleaning and verification software at this step in the process. The derived variable code included routines for calculating these variables, treating missing data appropriately, adding variable labels, etc. This code was based on the Main Survey (MS) Data Analysis Plan in which it was outlined that approximately 219 derived variables were calculated from PISA MS data.

Further explained in the MS Analysis Plan, for all questions in the MS questionnaires that were not converted into derived variables, the international database contained item-level data as obtained from the delivery platform. These included single-item constructs that could be measured without any transformation (e.g., ST002 Study program, ST016 Life satisfaction, ST021 Age of immigration, ST111 ISCED-level of educational aspiration, SC013 School type: public vs private, SC014 School management), as well as multi-item questions that were used by analysts for their respective needs (e.g., ST063 School science courses attended, ST064 Freedom of curricular choice, ST076/078 Activities before/after school, and most questions from the School Questionnaire). Derived variables were specified in line with previous cycles of PISA wherever possible. In terms of this alignment, first priority was given to alignment with PISA 2006, to enable comparison on science-related issues. Second priority was given to PISA 2012, to enable stability across recent and future cycles. For IRT scales, only alignment with PISA 2006 was included. See Chapter 16 for more information on derived variables.

As this phase of the processing was completed, all derivations were checked by DIPF. Any updates or recoding made to the derived variable code were completed and documented and redelivered to the Data Management team for use in the cleaning and verification software. Data files were refreshed appropriately with this new code to include all updates to these variables.

### **DELIVERABLES**

After all data processing steps were complete and all updates to the data were made by National Centres to resolve any issues or inconsistencies, the final phase of data processing included the creation of deliverable files for all core contractors as well as the National Centres. Each data file deliverable required a unique specification of variables along with their designated ordering within the file.

In addition to the generation of files for contractors and National Centre use, the ‘deliverables’ step in the cleaning and verification process contained critical applications to the data – such as the application of proxy scores, plausible values, background questionnaire scales, and weights. The dynamic feature of the cleaning and verification software allowed for the Data Management team to tailor specific deliverables.

ETS Data Management produced a database containing the PISA 2015 data for National Centres and provided specific deliverables for core contractors as well as the OECD Secretariat according to particular specifications. In order to produce these customised files for contractors, each deliverable required a separate series of checks and reviews in order to ensure all data were handled appropriately and all values were populated as expected.

### **Preparing files for public use and analysis**

In order to prepare for the public release of PISA 2015 main survey data, ETS Data Management provided data files in SPSS and SAS to National Centres and the OECD Secretariat in batch deliveries at various review points during the main survey cycle. With the initial data deliveries of the main survey, the data files included proxy proficiency scores



for analysis. These data were later updated to include plausible values and questionnaire indices. During each of these phases of delivery, National Centres reviewed these data files and provided ETS Data Management with any comments and/or revisions to the data.

### **Files prepared for national centre data reviews**

During the PISA 2015 main survey, the following files were prepared and released to National Centres at different stages of the data review:

- **Student combined data file** contained all student responses for test items (raw and scored), background questionnaire items, financial literacy items (if applicable), collaborative problem-solving items (if applicable), and optional questionnaire items such as Parent Questionnaire, Educational Career (EC) Questionnaire, Information and Computer Technology Literacy Familiarity (ICT) Questionnaire. These files included all raw variables, questionnaire indices, sampling weights, replicate weights, and plausible values.
- **School data file** contained all data from the School Questionnaires. These files included all raw variables, questionnaire indices, sampling weights, replicate weights, and plausible values.
- **Teacher data file** (if applicable) comprised data from the Teacher Questionnaire. These files included all raw variables, questionnaire indices and plausible values. In PISA 2015, Westat sampling did not calculate teacher weights and as such, there were no teacher weights in the data files.
- **Masked international database file** was a concatenated file of all countries provided further information for analysis to National Centres. In order to preserve country anonymity in this file, data files were ‘masked’ following specific guidelines from the OECD Secretariat that included issuing ‘alternate’ codes or required special handling for country identifiers.
- **Preliminary Public Use File** was produced toward the end of the PISA 2015 main survey and provided the National Centre with their country’s own data as it would be presented in the final public release. These data included all country-requested variable suppressions. More information on the suppression period is discussed later in this chapter.
- **Analysis Reports** were delivered by data management and analysis and used by contractors and National Centres for quality control and validation purposes: plausibility of 1) distributions of background characteristics and 2) performance results for groups, especially in the extent to which they agree with expectations or external/historical information. These reports included:
  - **BQ Crosstabs:** An Excel file with cross tabulations of numeric categorical variables from the country’s Background Questionnaire.
  - **BQ MSIGS:** An Excel file of summary statistics for all numerical variables from the country’s Background Questionnaire.
  - **BQ SDTs:** Sets of country files containing summary data tables that provided descriptive statistics for every categorical background variable in the respective country’s PISA data file. For each country, the summary data tables included both international and country-specific background variables.
  - **Item Analysis Reports:** The item analysis tables contained summary information about the response types given by the respondents to the cognitive items. They contained, for each country, the percent of individuals choosing each option for multiple-choice items or the percent of individuals receiving each score in the scoring guide for the constructed-response items. They also contained the international average percentages for each response category.

### **Records included in and excluded from the database**

The following records were included in the database:

- student files
  - all PISA student respondents who participated in either the paper-based or computer-based assessment
  - all PISA students who had any response data or who were part of the original country sample
- school files
  - all participating schools – specifically, any school with a student included in the PISA sample and with a record in the school-level international database regardless of whether or not the school returned the School Questionnaire
- Teacher files
  - all PISA teacher participants that were included in the original sample.

The following records were excluded from the database<sup>5</sup>:

- student files
  - additional data collected by countries as part of national options
  - students who did not have the minimum response data to be considered a “respondent”<sup>6</sup>
  - students who refused to participate in the assessment sessions
- school files
  - additional data collected by countries as part of national options
- teacher files
  - teachers who refused to participate in the questionnaire.

### Categorising missing data

Within the data files, the coding of the data distinguishes between four different types of missing data:

1. Missing/blank is used to indicate the respondent was not presented the question according to the survey design or ended the assessment early and did not see the question. In the questionnaire data, it is only used to indicate that the respondent ended the assessment early or despite the opportunity, did not take the questionnaire.
2. No response/Omit indicates that the respondent had an opportunity to answer the question but did not respond.
3. Invalid is used to indicate a questionnaire item was suppressed by country request or that an answer was not conforming to the expected response. For a paper-based questionnaire, the respondent indicated more than one choice for an exclusive-choice question. For a computer-based questionnaire, the response was not in an acceptable range of responses, e.g., the response to a question asking for a percentage was greater than 100.
4. Not applicable indicates that a response was provided even though the response to an earlier question should have directed the respondent to skip that question, or the response could not be determined due to a printing problem or torn booklet. In the questionnaire data, it is also used to indicate missing by design (i.e. the respondent was never given the opportunity to see this question).
5. Valid skip indicates that the question was not answered because a response to an earlier question directed the respondent to skip the question. This code was assigned during data processing.

### Data management and confidentiality, variable suppressions

During the PISA 2015 cycle, some country regulations and laws restricted the sharing of data, as originally collected, with other countries. The key goal of such disclosure control is to prevent the accidental or intentional identification of individuals in the release of data. However, suppression of information or reduction of detail clearly impacts the analytical utility of the data. Therefore, both goals must be carefully balanced. As a general directive for PISA 2015, the OECD requested that all countries make available the largest permissible set of information at the highest level of disaggregation possible.

Each country was required to provide early notification of any rules affecting the disclosure and sharing of PISA sampling, operational or response data. Furthermore, each country was responsible for implementing any additional confidentiality measures in the database before delivery to the Consortium. Most importantly, any confidentiality edits that changed the response values had to be applied prior to submitting data in order to work with identical values during processing, cleaning and analysis. The DME software only supported the suppression of entire variables. All other measures were implemented under the responsibility of the country via the export/import functionality or by editing individual data cells.

With the delivery of the data from the National Centre, the Data Management team reviewed a detailed document of information that included any implemented or required confidentiality practices in order to evaluate the impact on the data management cleaning and analysis processes. Country suppression requests generally involved specific variables that violate confidentiality and anonymity of student, school, and/or teacher data, as well as technical errors in the data that could not be resolved through contractor cleaning and verification procedures. A listing of suppressions at the country variable-level is located in Figure 10.7 at the end of this chapter.



■ Figure 10.6 [Part 1/3] ■

**PISA 2015 range restriction rules for inconsistent and extreme values for main survey data**

Sequence	Dataset (STU, SCH, TCH)	Description of rule	SAS code
<b>Student dataset</b>			
1	STU	Invalidate if number of class periods per week in test language (ST059Q01TA) is greater than 40.	if ( ST059Q01TA > 40) then ST059Q01TA =.l;
2	STU	Invalidate if number of class periods per week in maths (ST059Q02TA) is greater than 40.	if ( ST059Q02TA > 40) then ST059Q02TA =.l;
3	STU	Invalidate if number of class periods per week in science (ST059Q03TA) is greater than 40.	if ( ST059Q03TA > 40) then ST059Q03TA =.l;
4	STU	Invalidate if number of total class periods in a week (ST060Q01NA) is greater than 120.	if (ST060Q01NA > 120) then ST060Q01NA =.l;
5	STU	Invalidate if average number of minutes in a class period (ST061Q01NA) is less than 10 or greater than 120.	if (ST061Q01NA > 120 or ST061Q01NA < 10) then ST061Q01NA =.l;
6	STU	Invalidate if age of child starting ISCED 1 (PA014Q01NA) is greater than 14.	if PA014Q01NA > 14 then PA014Q01NA =.l;
7	STU	Invalidate if repeated a grade in ISCED3 (ST127Q03TA) but currently in ISCED2.	if ISCEDL = 2 then ST127Q03TA =.l;
8	STU	Mark as missing if learning time per week in maths (MMINS) is greater than 2400 min (40 hours).	if MMINS > 2400 then MMINS =.M;
9	STU	Mark as missing if learning time per week in test language (LMINS) is greater than 2400 min (40 hours).	if LMINS > 2400 then LMINS =.M;
10	STU	Mark as missing if learning time per week in science (SMINS) is greater than 2400 min (40 hours).	if SMINS > 2400 then SMINS =.M;
11	STU	Mark as missing if learning time per week in total (TMINS) is greater than 3000 min (50 hours) or less than the sum of the parts (MMINS, LMINS, SMINS).	if TMINS > 3000 then TMINS =.M; if TMINS < sum(LMINS, MMINS, SMINS) then TMINS =.M;
12	STU	Mark as missing if out-of-school study time per week (OUTHOURS) is greater than 70 hours.	if OUTHOURS > 70 then OUTHOURS = .M;
13	STU	Invalidate if age started ISCED 1 is greater than 16 or less than 2.	if (ST126Q02TA > 16 or ST126Q02TA < 2) then ST126Q02TA =.l;
<b>School dataset</b>			
1	SCH	Invalidate if number of computers connected to the internet (SC004Q03TA) is greater than the number of computers available to students (SC004Q02TA).	if SC004Q03TA > SC004Q02TA then SC004Q03TA =.l;
2	SCH	Invalidate if number of portable computers (SC004Q04NA) is greater than the number of computers available to students (SC004Q02TA).	if SC004Q04NA > SC004Q02TA then SC004Q04NA =.l;
3	SCH	Invalidate if total number of full time teachers (SC018Q01TA01) is negative.	if (SC018Q01TA01 < 0) then SC018Q01TA01 =.l;
4	SCH	Invalidate if number of full time certified teachers (SC018Q02TA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q02TA01 > SC018Q01TA01 then SC018Q02TA01 =.l;
5	SCH	Invalidate if number of full time Bachelor degree teachers (SC018Q05NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q05NA01 > SC018Q01TA01 then SC018Q05NA01 =.l;
6	SCH	Invalidate if number of full time Master's degree teachers (SC018Q06NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q06NA01 > SC018Q01TA01 then SC018Q06NA01 =.l;
7	SCH	Invalidate if number of full time ISCED 6 teachers (SC018Q07NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC018Q07NA01 > SC018Q01TA01 then SC018Q07NA01 =.l;
8	SCH	Invalidate if number of part time certified teachers (SC018Q02TA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q02TA02 > SC018Q01TA02 then SC018Q02TA02 =.l;
9	SCH	Invalidate if number of part time Bachelor degree teachers (SC018Q05NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q05NA02 > SC018Q01TA02 then SC018Q05NA02 =.l;
10	SCH	Invalidate if number of part time Master's degree teachers (SC018Q06NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q06NA02 > SC018Q01TA02 then SC018Q06NA02 =.l;
11	SCH	Invalidate if number of part time ISCED 6 teachers (SC018Q07NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC018Q07NA02 > SC018Q01TA02 then SC018Q07NA02 =.l;
12	SCH	Invalidate if total number of full time science teachers (SC019Q01NA01) is negative.	if (SC019Q01NA01 < 0) then SC019Q01NA01 =.l;
13	SCH	Invalidate if number of full time science teachers (SC019Q01NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC019Q01NA01 > SC018Q01TA01 then SC019Q01NA01 =.l;

■ Figure 10.6 [Part 2/3] ■

**PISA 2015 range restriction rules for inconsistent and extreme values for main survey data**

Sequence	Dataset (STU, SCH, TCH)	Description of rule	SAS code
14	SCH	Invalidate if number of full time certified science teachers (SC019Q02NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC019Q02NA01 > SC018Q01TA01 then SC019Q02NA01 =.l;
15	SCH	Invalidate if number of full time ISCED 5A science teachers (SC019Q03NA01) exceeds total number of full time teachers (SC018Q01TA01).	if SC019Q03NA01 > SC018Q01TA01 then SC019Q03NA01 =.l;
16	SCH	Invalidate if number of part time science teachers (SC019Q01NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC019Q01NA02 > SC018Q01TA02 then SC019Q01NA02 =.l;
17	SCH	Invalidate if number of part time certified science teachers (SC019Q02NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC019Q02NA02 > SC018Q01TA02 then SC019Q02NA02 =.l;
18	SCH	Invalidate if number of part time ISCED 5A science teachers (SC019Q03NA02) exceeds total number of part time teachers (SC018Q01TA02).	if SC019Q03NA02 > SC018Q01TA02 then SC019Q03NA02 =.l;
19	SCH	Invalidate if number of full time certified science teachers (SC019Q02NA01) exceeds total number of full time science teachers (SC019Q01NA01).	if SC019Q02NA01 > SC019Q01NA01 then SC019Q02NA01 =.l;
20	SCH	Invalidate if number of full time ISCED 5A science teachers (SC019Q03NA01) exceeds total number of full time science teachers (SC019Q01NA01).	if SC019Q03NA01 > SC019Q01NA01 then SC019Q03NA01 =.l;
21	SCH	Invalidate if number of part time certified science teachers (SC019Q02NA02) exceeds total number of part time science teachers (SC019Q01NA02).	if SC019Q02NA02 > SC019Q01NA02 then SC019Q02NA02 =.l;
22	SCH	Invalidate if number of part time ISCED 5A science teachers (SC019Q03NA02) exceeds total number of part time science teachers (SC019Q01NA02).	if SC019Q03NA02 > SC019Q01NA02 then SC019Q03NA02 =.l;
23	SCH	Invalidate if sum of funding percentages is less than 98% or greater than 102% (SC016Q01TA + SC016Q02TA + SC016Q03TA + SC016Q04TA).	if sum(SC016Q01TA, SC016Q02TA, SC016Q03TA, SC016Q04TA) > 102 or sum (SC016Q01TA, SC016Q02TA, SC016Q03TA, SC016Q04TA) < 98 then do; SC016Q01TA =.l; SC016Q02TA =.l; SC016Q03TA =.l; SC016Q04TA =.l; end;
24	SCH	Invalidate if percentage of teaching staff (SC025Q01NA) is greater than 100%.	if SC025Q01NA > 100 then SC025Q01NA =.l;
25	SCH	Invalidate if percentage of science teacher staff (SC025Q02NA) is greater than 100%.	if SC025Q02NA > 100 then SC025Q02NA =.l;
26	SCH	Invalidate if percentage of students with <heritage language> different than <test language> (SC048Q01NA) is greater than 100%.	if SC048Q01NA > 100 then SC048Q01NA =.l;
27	SCH	Invalidate if percentage of students with special needs (SC048Q02NA) is greater than 100%.	if SC048Q02NA > 100 then SC048Q02NA =.l;
28	SCH	Invalidate if percentage of students from disadvantaged homes (SC048Q03NA) is greater than 100%.	if SC048Q03NA > 100 then SC048Q03NA =.l;
29	SCH	Invalidate if percentage of parents that initiated discussion on child (SC064Q01TA) is greater than 100%.	if SC064Q01TA > 100 then SC064Q01TA =.l;
30	SCH	Invalidate if percentage of parents where teacher initiated discussion on child (SC064Q02TA) is greater than 100%.	if SC064Q02TA > 100 then SC064Q02TA =.l;
31	SCH	Invalidate if percentage of parents participated in school government (SC064Q03TA) is greater than 100%.	if SC064Q03TA > 100 then SC064Q03TA =.l;
32	SCH	Invalidate if percentage of parents that volunteered in extracurricular activities (SC064Q04NA) is greater than 100%.	if SC064Q04NA > 100 then SC064Q04NA =.l;
33	SCH	Invalidate if total number of boys (SC002Q01TA) and total number of girls (SC002Q02TA) are both zero.	if SC002Q01TA = 0 and SC002Q02TA = 0 then do; SC002Q01TA =.l; SC002Q02TA =.l; end;
34	SCH	Invalidate if total number of students in modal grade (SC004Q01TA) is greater than total number of students (SC002Q01TA + SC002Q02TA).	if SC004Q01TA > sum(SC002Q01TA,SC002Q02TA) then SC004Q01TA =.l;
35	SCH	Invalidate if total number of part time teachers (SC018Q01TA02) is negative.	if SC018Q01TA02 < 0 then SC018Q01TA02 =.l;
36	SCH	Mark index of computer availability (RATCMP1) as missing if there are only 10 or less students in the modal grade.	If SC004Q01TA <= 10 then RATCMP1 =.M;
37	SCH	Mark index of computers connected to the Internet (RATCMP2) as missing if there are only 10 or less students in the modal grade.	If SC004Q01TA <= 10 then RATCMP2 =.M;
38	SCH	Recode student-teacher ratio (STRATIO) to set the minimum number of teachers at 1 and then to set the final ratio to a maximum of 100 and a minimum of 1.	if nmiss(SCHSIZE,TOTAT) = 0 then STRATIO = max(min(SCHSIZE/max(1, TOTAT), 100), 1); else STRATIO =.M;



■ Figure 10.6 [Part 3/3] ■

**PISA 2015 range restriction rules for inconsistent and extreme values for main survey data**

Sequence	Dataset (STU, SCH, TCH)	Description of rule	SAS code
<b>Teacher dataset</b>			
1	TCH	Invalidate if number of years teaching at school (TC007Q01NA) exceeds reported age (TC002Q01NA) minus 15.	if TC007Q01NA > (TC002Q01NA - 15) then TC007Q01NA =.;
2	TCH	Invalidate if total number of years teaching (TC007Q02NA) exceeds reported age (TC002Q01NA) minus 15.	if TC007Q02NA > (TC002Q01NA - 15) then TC007Q02NA =.;
3	TCH	Invalidate if years working as a teacher in total (TC007Q02NA) is less than years working as a teacher in this school (TC007Q01NA).	if TC007Q01NA > TC007Q02NA then TC007Q01NA =.;
4	TCH	Invalidate if proportion of teacher education dedicated to <broad science> and technology content (TC029Q01NA) + <school science> (TC029Q02NA) + general topics (TC029Q03NA) + other topics (TC029Q04NA) is less than 98% or greater than 102%.	if sum(TC029Q01NA, TC029Q02NA, TC029Q03NA, TC029Q04NA) > 102 or sum(TC029Q01NA, TC029Q02NA, TC029Q03NA, TC029Q04NA) < 98 then do; TC029Q01NA =.; TC029Q02NA =.; TC029Q03NA =.; TC029Q04NA =.; end;
5	TCH	Invalidate if proportion of professional development activities dedicated to <broad science> and technology content (TC030Q01NA) + <school science> (TC030Q02NA) + general topics (TC030Q03NA) + other topics (TC030Q04NA) is less than 98% or greater than 102%.	if sum(TC030Q01NA, TC030Q02NA, TC030Q03NA, TC030Q04NA) > 102 or sum(TC030Q01NA, TC030Q02NA, TC030Q03NA, TC030Q04NA) < 98 then do; TC030Q01NA =.; TC030Q02NA =.; TC030Q03NA =.; TC030Q04NA =.; end;

■ Figure 10.7 [Part 1/2] ■

**PISA 2015 main survey country/variable suppression list**

Country	Variable
AUT	Stratum SC002Q01TA, SC002Q02TA, SCHSIZE
AUS	Student financial literacy data
BEL (Flemish only)	SC013Q01TA, SC014Q01NA, SC016Q01TA, SC016Q02TA, SC016Q03TA, SC016Q04TA, SCHLTYPE
QCH	Stratum, Region
DEU	STRATUM
ISR	SC013, SC014, SC016, SCHLTYPE, STRATUM
ITA	STRATUM
CYP <sup>1,2</sup>	STRATUM, LANGTEST_COG, LANGTEST_QQQ, LANGTEST, SC001Q01TA
KAZ	STRATUM
NZL	SC002Q01TA, SC002Q02TA, SC004Q01TA, SC004Q02TA, SC014Q01NA, SCHSIZE
QMA, QNC <sup>3</sup> and PRI <sup>4</sup> (USA)	All school variables, All teacher variables, CNTSCHID <sup>5</sup> , ST001D01T, ST003D02T, ST003D03T, ST005Q01TA, ST006Q01TA, ST006Q02TA, ST006Q03TA, ST006Q04TA, ST007Q01TA, ST008Q01TA, ST008Q02TA, ST008Q03TA, ST008Q04TA, ST019AQ01T, ST019BQ01T, ST019CQ01T, ST021Q01TA, ST022Q01TA, AGE, ISCEDL, ISCEDD, ISCEDO, GRADE, IMMIG, MISCED, FISCED, HISCED, BMMJ1, BFMJ2, HISEI, PARED, COBN_F, COBN_M, COBN_S, LANGN, OCOD1, OCOD2, UNIT, WVARSTRR
SVN	ST063, ST064, ST065, ST103, ST104, ST107, TDTEACH, PERFEED, ADINST
SWE	ST003D02T, ST003D03T SC001Q01TA, SC002Q01TA, SC002Q02TA, SC003Q01TA, SC004Q01TA, SC013Q01TA, SC014Q01NA, SC016Q01TA, SC016Q02TA, SC016Q03TA, SC016Q04TA, SC018Q01TA01 SC018Q01TA02, SC018Q02TA01 SC018Q02TA02, SC018Q05NA01 SC018Q05NA02, SC018Q06NA01 SC018Q06NA02, SC018Q07NA01 SC018Q07NA02, SC019Q01NA01 SC019Q01NA02, SC019Q02NA01 SC019Q02NA02, SC019Q03NA01 SC019Q03NA02, SC048Q01NA SC048Q02NA, SC048Q03NA

■ Figure 10.7 [Part 2/2] ■

**PISA 2015 main survey country/variable suppression list**

Country	Variable
TAP	STRATUM
THA	STRATUM

1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

2. CYP data is suppressed from the public use files. Variables were suppressed in the national data files.

3. QMA (Massachusetts) and QNC (North Carolina) are the United States state samples analysed in the PISA 2015 main survey.

4. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

5. With this suppression request, all school and teacher data is suppressed. As a result, CNTSCHID is suppressed in all data files.

**Notes**

1. "Data Management" refers to the collective set of activities and tasks that each country had to perform to produce the required national database.

2. Teachers who were absent, excluded, or refused to participate in the session may be marked as a "non-participant."

3. For example, if an inconsistency existed between age and seniority, the proposed rules invalidates seniority but keeps "age".

4. With this principle, the original values were kept, while the values for the derived variable may have the applied "invalid" rule.

5. Due to issues identified during data adjudication, data from Argentina, Kazakhstan, Malaysia and Albania, student questionnaire data (only) have been extracted into a separate file for analysis.

6. To be considered a "respondent" the student must have one test item response and a minimum number of responses to the student background questionnaire (that included responses for ST012 or ST013); or, responded to at least half of the number of test items in his or her booklet/form.



---

**11**

## Sampling outcomes

<b>Population coverage.....</b>	204
<b>School and student response rates.....</b>	205
<b>Teacher response rates.....</b>	214
<b>Design effects and effective sample sizes .....</b>	215
<b>Variability of the design effect .....</b>	217

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

This chapter reports on PISA sampling outcomes. Details of the sample design are provided in Chapter 4.

## POPULATION COVERAGE

Tables 11.1 and 11.2 (by adjudicated regions) show the quality indicators for population coverage and the information used to develop them. The following notes explain the meaning of each coverage index and how the data in each column of the table were used.

Coverage indices 1, 2 and 3 are intended to measure PISA population coverage. Coverage indices 4 and 5 are intended to be diagnostic in cases where indices 1, 2 or 3 have unexpected values. Many references are made in this chapter to the various sampling tasks on which National Project Managers (NPMs) documented statistics and other information needed in undertaking the sampling of schools and students. Note that although no comparison is made between the total population of 15-year-olds and the enrolled population of 15-year-old students, generally the enrolled population was expected to be less than or equal to the total population. Occasionally this was not the case due to differing data sources for these two values.

Coverage index 1: Coverage of the national population, calculated by  $P/(P + E) \times (ST7b\_3/ST7b\_1)$ :

- Coverage index 1 shows the extent to which the weighted participants covered the final target population after all school exclusions. The following bullet points give details of its computation.
  - In the preceding expression  $P/(P + E)$  broadly represents the coverage proportion due to within-school exclusion, and  $(ST7b\_3/ST7b\_1)$  the coverage proportion due to school-level exclusion.
  - The national population value, defined by sampling task 7b response box [1] and denoted here as  $ST7b\_1$  (and in Table 11.1 as the target population) is the population that includes all enrolled 15-year-old students in grades 7 and above in each participating country (with the possibility of small levels of exclusions), based on national statistics. However, the final national population value reflected for each country's school sampling frame might have had some school-level exclusions. The value that represents the population of enrolled 15-year-old students minus those in excluded schools is represented initially by response box [3] on sampling task 7b. It is denoted here as  $ST7b\_3$ . As in PISA 2012, the procedure for PISA 2015 was that small schools having only one or two PISA-eligible students could not be excluded from the school frame but could be excluded in the field if the school still had only one or two PISA-eligible students at the time of data collection. Therefore, what is noted in coverage index 1 as  $ST7b\_3$  (and in Table 11.1 as target minus school-level exclusions) was a number after accounting for all school-level exclusions, which means a number that omits schools excluded from the sampling frame in addition to those schools excluded in the field. Thus, the term  $(ST7b\_3/ST7b\_1)$  provides the proportion of the national population covered in each country based on national statistics.
  - The value  $(P + E)$  provides the weighted estimate from the student sample of all PISA-eligible 15-year-olds in each participating country, where  $P$  is the weighted estimate of PISA-eligible non-excluded 15-year-old students and  $E$  is the weighted estimate of PISA-eligible 15-year-old students that were excluded within schools. Therefore, the term  $P/(P + E)$  provides an estimate, based on the student sample, of the proportion of the PISA-eligible 15-year-old population represented by the non-excluded PISA-eligible 15-year-old students.
  - The result of multiplying these two proportions together  $P/(P + E)$  and  $(ST7b\_3/ST7b\_1)$  indicates the overall proportion of the national population covered by the non-excluded portion of the student sample.

Coverage index 2: Coverage of the national enrolled population, calculated by  $P/(P + E) \times (ST7b\_3/ST7a\_2.1)$ :

- Coverage index 2 shows the extent to which the weighted participants covered the target population of all enrolled students in grades 7 and above.
- The national enrolled population (NEP), defined by sampling task 7a response box [2.1] and denoted here as  $ST7a\_2.1$  (and as enrolled 15-year-old students in Table 11.1), is the population that includes all enrolled 15-year-old students in grades 7 and above in each participating country, based on national statistics. The final national population, denoted here as  $ST7b\_3$  as described above for coverage index 1, reflects the 15-year-old population after school-level and other small exclusions. This value represents the population of enrolled 15-year-old students less those in excluded schools.
- The value  $(P + E)$  provides the weighted estimate from the student sample of all eligible 15-year-olds in each country, where  $P$  is the weighted estimate of PISA-eligible non-excluded 15-year-old students and  $E$  is the weighted estimate of PISA-eligible 15-year-old students that were excluded within schools. Therefore, the term  $P/(P + E)$  provides an



estimate based on the student sample of the proportion of the PISA-eligible 15-year-old population that is represented by the non-excluded PISA-eligible 15-year-old students.

- Multiplying these two proportions together ( $P/(P + E)$  and  $(ST7b\_3/ST7a\_2.1)$ ) gives the overall proportion of the NEP that was covered by the non-excluded portion of the student sample.

Coverage index 1 and coverage index 2 will differ when countries have excluded geographical areas or language groups apart from other school-level exclusions. In these cases coverage index 2 will be less than coverage index 1.

Coverage index 3: Coverage of the national 15-year-old population, calculated by  $P/ST7a\_1$ :

- The national population of 15-year-olds, defined by sampling task 7a response box [1] and denoted here as  $ST7a\_1$  (and called all 15-year-olds in Table 11.1), is the entire population of 15-year-olds in each country (enrolled and not enrolled), based on national statistics. The value  $P$  is the weighted estimate of PISA-eligible non-excluded 15-year-old students from the student sample. Thus  $(P/ST7a\_1)$  indicates the proportion of the national population of 15-year-olds covered by the non-excluded portion of the student sample. It therefore also reflects the proportion of 15-year-olds excluded or not at school.

Coverage index 4: Coverage of the estimated school population, calculated by  $(P + E)/S$ :

- The value  $(P + E)$  provides the weighted estimate from the student sample of all PISA-eligible 15-year-old students in each country, where  $P$  is the weighted estimate of PISA-eligible non-excluded 15-year-old students and  $E$  is the weighted estimate of PISA-eligible 15-year-old students who were excluded within schools.
- The value  $S$  is an estimate of the 15-year-old school population in each participating country (called estimate of enrolled students from frame in Table 11.1). This is based on the actual or (more often) approximate number of 15-year-old students enrolled in each school in the sample, prior to contacting the school to conduct the assessment. The  $S$  value is calculated as the sum over all sampled schools of the product of each school's sampling weight and its number of 15-year-old students ( $ENR$ ) as recorded on the school sampling frame.
- Thus,  $(P + E)/S$  is the proportion of the estimated school 15-year-old population that is represented by the weighted estimate from the student sample of all PISA-eligible 15-year-old students. It is influenced by the accuracy of the school sample frame, fluctuations in the target population size and the accuracy of the within-school sampling process. Its purpose is to check whether the student sampling has been carried out correctly, and to assess whether the value of  $S$  is a reliable measure of the number of enrolled 15-year-olds. This is important for interpreting coverage index 5.

Coverage index 5: Coverage of the school sampling frame population, calculated by  $S/ST7b\_3$ :

- The value  $(S/ST7b\_3)$  is the ratio of the enrolled 15-year-old population, as estimated from data on the school sampling frame, to the size of the enrolled student population, as reported on sampling task 7b and adjusted by removing any additional excluded schools in the field. In some cases, this provided a check as to whether the data on the sampling frame gave a reliable estimate of the number of 15-year-old students in each school. In other cases, however, it was evident that  $ST7b\_3$  had been derived using data from the sampling frame by the NPM, so that this ratio may have been close to 1.0 even if enrolment data on the school sampling frame were poor. Under such circumstances, coverage index 4 would differ noticeably from 1.0, and the figure for  $ST7b\_3$  would also be inaccurate.

## SCHOOL AND STUDENT RESPONSE RATES

Tables 11.3 to 11.8 present school and student-level response rates at the national and regional levels.

- Tables 11.3 and 11.4 (by adjudicated regions) indicate the rates calculated by using only original schools and no replacement schools.
- Tables 11.5 and 11.6 (by adjudicated regions) indicate the improved response rates when first and second replacement schools were accounted for in the rates.
- Tables 11.7 and 11.8 (by adjudicated regions) indicate the student response rates among the full set of participating schools.

[Part 1/2]

Table 11.1 PISA target populations and samples

	All 15-year-olds	Enrolled 15-year-olds	Target population	School-level exclusions	Target minus school level exclusions	School level exclusion rate (%)	Estimation of enrolled students from frame	Number of participating students	Weighted number of participating students	Number of excluded students
<b>OECD</b>										
Australia	282 888	282 547	282 547	6 940	275 607	2.46	276 072	14 530	256 329	681
Austria	88 013	82 683	82 683	790	81 893	0.96	81 730	7 007	73 379	84
Belgium	123 630	121 954	121 694	1 597	120 097	1.31	118 915	9 651	114 902	39
Canada	396 966	381 660	376 994	1 590	375 404	0.42	381 133	20 058	331 546	1 830
Chile	255 440	245 947	245 852	2 641	243 211	1.07	232 756	7 053	203 782	37
Czech Republic	90 391	90 076	90 076	1 814	88 262	2.01	87 999	6 894	84 519	25
Denmark	68 174	67 466	67 466	605	66 861	0.90	63 897	7 161	60 655	514
Estonia	11 676	11 491	11 491	416	11 075	3.62	11 154	5 587	10 834	116
Finland	58 526	58 955	58 955	472	58 483	0.80	58 782	5 882	56 934	124
France	807 867	778 679	778 679	28 742	749 937	3.69	749 284	6 108	734 944	35
Germany	774 149	774 149	774 149	11 150	762 999	1.44	794 206	6 522	743 969	54
Greece	105 530	105 253	105 253	953	104 300	0.91	103 031	5 532	96 157	58
Hungary	94 515	90 065	90 065	1 945	88 120	2.16	89 808	5 658	84 644	55
Iceland	4 250	4 195	4 195	17	4 178	0.41	4 163	3 374	3 966	131
Ireland	61 234	59 811	59 811	72	59 739	0.12	61 461	5 741	59 082	197
Israel	124 852	118 997	118 997	2 310	116 687	1.94	115 717	6 598	117 031	115
Italy	616 761	567 268	567 268	11 190	556 078	1.97	516 113	11 583	495 093	246
Japan	1 201 615	1 175 907	1 175 907	27 323	1 148 584	2.32	1 151 305	6 647	1 138 349	2
Korea	620 687	619 950	619 950	3 555	616 395	0.57	615 107	5 581	569 106	20
Latvia	17 255	16 955	16 955	677	16 278	3.99	16 334	4 869	15 320	70
Luxembourg	6 327	6 053	6 053	162	5 891	2.68	5 891	5 299	5 540	331
Mexico	2 257 399	1 401 247	1 401 247	5 905	1 395 342	0.42	1 373 919	7 568	1 392 995	30
Netherlands	201 670	200 976	200 976	6 866	194 110	3.42	191 966	5 385	191 817	14
New Zealand	60 162	57 448	57 448	681	56 767	1.19	56 875	4 520	54 274	333
Norway	63 642	63 491	63 491	854	62 637	1.35	61 809	5 456	58 083	345
Poland	380 366	361 600	361 600	6 122	355 478	1.69	355 158	4 478	345 709	34
Portugal	110 939	101 107	101 107	424	100 683	0.42	102 193	7 325	97 214	105
Slovak Republic	55 674	55 203	55 203	1 376	53 827	2.49	54 499	6 350	49 654	114
Slovenia	18 078	17 689	17 689	290	17 399	1.64	17 286	6 406	16 773	114
Spain	440 084	414 276	414 276	2 175	412 101	0.53	409 246	6 736	399 935	200
Sweden	97 749	97 210	97 210	1 214	95 996	1.25	94 097	5 458	91 491	275
Switzerland	85 495	83 655	83 655	2 320	81 335	2.77	81 026	5 860	82 223	107
Turkey	1 324 089	1 100 074	1 100 074	5 746	1 094 328	0.52	1 091 317	5 895	925 366	31
United Kingdom	747 593	746 328	746 328	23 412	722 916	3.14	707 415	14 157	627 703	870
United States	4 220 325	3 992 053	3 992 053	12 001	3 980 052	0.30	3 902 089	5 712	3 524 497	193
<b>Partners</b>										
Albania	48 610	45 163	45 163	10	45 153	0.02	43 919	5 215	40 896	0
Algeria	389 315	354 936	354 936	354 936	354 936	0.00	355 216	5 519	306 647	0
Argentina	718 635	578 308	578 308	2 617	575 691	0.45	572 941	6 349	394 917	21
Brazil	3 803 681	2 853 388	2 853 388	64 392	2 788 996	2.26	2 692 686	23 141	2 425 961	119
B-S-J-G (China)*	2 084 958	1 507 518	1 507 518	58 639	1 448 879	3.89	1 437 201	9 841	1 331 794	33
Bulgaria	66 601	59 397	59 397	1 124	58 273	1.89	56 483	5 928	53 685	49
Colombia	760 919	674 079	674 079	37	674 042	0.01	673 817	11 795	567 848	9
Costa Rica	81 773	66 524	66 524	66 524	66 524	0.00	67 073	6 866	51 897	13
Croatia	45 031	35 920	35 920	805	35 115	2.24	34 652	5 809	40 899	86
Cyprus <sup>1</sup>	9 255	9 255	9 253	109	9 144	1.18	9 126	5 571	8 785	228
Dominican Republic	193 153	139 555	139 555	2 382	137 173	1.71	138 187	4 740	132 300	4
FYROM	16 719	16 717	16 717	259	16 458	1.55	16 472	5 324	15 847	8
Georgia	48 695	43 197	43 197	1 675	41 522	3.88	41 595	5 316	38 334	35
Hong Kong (China)	65 100	61 630	61 630	708	60 922	1.15	60 716	5 359	57 662	36
Indonesia	4 534 216	3 182 816	3 182 816	4 046	3 178 770	0.13	3 176 076	6 513	3 092 773	0
Jordan	126 399	121 729	121 729	71	121 658	0.06	119 024	7 267	108 669	70
Kazakhstan	211 407	209 555	209 555	7 475	202 080	3.57	202 701	7 841	192 909	0
Kosovo	31 546	28 229	28 229	1 156	27 073	4.10	26 924	4 826	22 333	50
Lebanon	64 044	62 281	62 281	1 300	60 981	2.09	60 882	4 546	42 331	0
Lithuania	33 163	32 097	32 097	573	31 524	1.79	31 588	6 525	29 915	227
Macao (China)	5 100	4 417	4 417	3	4 414	0.07	4 414	4 476	4 507	0
Malaysia	540 000	448 838	448 838	2 418	446 420	0.54	446 237	8 861	412 524	41
Malta	4 397	4 406	4 406	63	4 343	1.43	4 343	3 634	4 296	41
Moldova	31 576	30 601	30 601	182	30 419	0.59	30 145	5 325	29 341	21
Montenegro	7 524	7 506	7 506	40	7 466	0.53	7 312	5 665	6 777	300
Peru	580 371	478 229	478 229	6 355	471 874	1.33	470 651	6 971	431 738	13
Qatar	13 871	13 850	13 850	380	13 470	2.74	13 470	12 083	12 951	193
Romania	176 334	176 334	176 334	1 823	174 511	1.03	172 652	4 876	164 216	3
Russian Federation	1 176 473	1 172 943	1 172 943	24 217	1 148 726	2.06	1 189 441	6 036	1 120 932	13
Singapore	48 218	47 050	47 050	445	46 605	0.95	46 620	6 115	46 224	25
Chinese Taipei	295 056	287 783	287 783	1 179	286 604	0.41	286 778	7 708	251 424	22
Thailand	895 513	756 917	756 917	9 646	747 271	1.27	751 010	8 249	634 795	22
Trinidad and Tobago	17 371	17 371	17 371	17 371	17 371	0.00	17 371	4 692	13 197	0
Tunisia	122 186	122 186	122 186	679	121 507	0.56	122 767	5 375	113 599	3
United Arab Emirates	51 687	51 518	51 499	994	50 505	1.93	50 060	14 167	46 950	63
Uruguay	53 533	43 865	43 865	4	43 861	0.01	43 737	6 062	38 287	6
Viet Nam	1 803 552	1 032 599	1 032 599	6 557	1 026 042	0.63	996 757	5 826	874 859	0

\* B-S-J-G (China) refers to the four PISA-participating China provinces: Beijing, Shanghai, Jiangsu and Guangdong.

1. Note to Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.



[Part 2/2]  
Table 11.1 PISA target populations and samples

	Weighted number of excluded students	Number of ineligible students	Weighted number of ineligible students	Within-school exclusion rate (%)	Overall exclusion rate (%)	Percentage of ineligible / withdrawn	Coverage Index 1	Coverage Index 2	Coverage Index 3	Coverage Index 4	Coverage Index 5
<b>OECD</b>											
Australia	7 736	904	8 203	2.93	5.31	3.11	0.95	0.95	0.91	0.96	1.00
Austria	866	669	3 431	1.17	2.11	4.62	0.98	0.98	0.83	0.91	1.00
Belgium	410	147	1 576	0.36	1.66	1.37	0.98	0.98	0.93	0.97	0.99
Canada	25 340	864	9 513	7.10	7.49	2.67	0.93	0.91	0.84	0.94	1.02
Chile	1 393	114	3 782	0.68	1.75	1.84	0.98	0.98	0.80	0.88	0.96
Czech Republic	368	82	825	0.43	2.44	0.97	0.98	0.98	0.94	0.96	1.00
Denmark	2 644	48	289	4.18	5.04	0.46	0.95	0.95	0.89	0.99	0.96
Estonia	218	34	61	1.97	5.52	0.55	0.94	0.94	0.93	0.99	1.01
Finland	1 157	13	124	1.99	2.78	0.21	0.97	0.97	0.97	0.99	1.01
France	3 620	157	16 455	0.49	4.16	2.23	0.96	0.96	0.91	0.99	1.00
Germany	5 342	110	11 334	0.71	2.14	1.51	0.98	0.98	0.96	0.94	1.04
Greece	965	87	1 616	0.99	1.89	1.66	0.98	0.98	0.91	0.94	0.99
Hungary	1 009	48	769	1.18	3.31	0.90	0.97	0.97	0.90	0.95	1.02
Iceland	132	179	181	3.23	3.62	4.40	0.96	0.96	0.93	0.98	1.00
Ireland	1 825	117	1 033	3.00	3.11	1.70	0.97	0.97	0.96	0.99	1.03
Israel	1 803	78	1 323	1.52	3.43	1.11	0.97	0.97	0.94	1.03	0.99
Italy	9 395	305	11 766	1.86	3.80	2.33	0.96	0.96	0.80	0.98	0.93
Japan	318	12	1 868	0.03	2.35	0.16	0.98	0.98	0.95	0.99	1.00
Korea	1 806	65	6 268	0.32	0.89	1.10	0.99	0.99	0.92	0.93	1.00
Latvia	174	153	430	1.12	5.07	2.77	0.95	0.95	0.89	0.95	1.00
Luxembourg	331	24	24	5.64	8.16	0.41	0.92	0.92	0.88	1.00	1.00
Mexico	6 810	505	84 669	0.49	0.91	6.05	0.99	0.99	0.62	1.02	0.98
Netherlands	502	20	592	0.26	3.67	0.31	0.96	0.96	0.95	1.00	0.99
New Zealand	3 112	114	1 102	5.42	6.54	1.92	0.93	0.93	0.90	1.01	1.00
Norway	3 366	43	445	5.48	6.75	0.72	0.93	0.93	0.91	0.99	0.99
Poland	2 418	22	1 505	0.69	2.38	0.43	0.98	0.98	0.91	0.98	1.00
Portugal	860	239	2 699	0.88	1.29	2.75	0.99	0.99	0.88	0.96	1.01
Slovak Republic	912	130	999	1.80	4.25	1.98	0.96	0.96	0.89	0.93	1.01
Slovenia	247	75	144	1.45	3.07	0.84	0.97	0.97	0.93	0.98	0.99
Spain	10 893	45	2 366	2.65	3.16	0.58	0.97	0.97	0.91	1.00	0.99
Sweden	4 324	46	715	4.51	5.71	0.75	0.94	0.94	0.94	1.02	0.98
Switzerland	1 357	146	1 659	1.62	4.35	1.99	0.96	0.96	0.96	1.03	1.00
Turkey	5 359	533	73 779	0.58	1.10	7.93	0.99	0.99	0.70	0.85	1.00
United Kingdom	34 747	297	8 914	5.25	8.22	1.35	0.92	0.92	0.84	0.94	0.98
United States	109 580	330	191 378	3.02	3.31	5.27	0.97	0.97	0.84	0.93	0.98
<b>Partners</b>											
Albania	0	0	0	0.00	0.02	0.00	1.00	1.00	0.84	0.93	0.97
Algeria	0	0	0	0.00	0.00	0.00	1.00	1.00	0.79	0.86	1.00
Argentina	1 367	204	11 847	0.34	0.80	2.99	0.99	0.99	0.55	0.69	1.00
Brazil	13 543	1 582	143 969	0.56	2.80	5.90	0.97	0.97	0.64	0.91	0.97
B-S-J-G (China)	3 609	552	94 478	0.27	4.15	7.07	0.96	0.96	0.64	0.93	0.99
Bulgaria	433	74	681	0.80	2.68	1.26	0.97	0.97	0.81	0.96	0.97
Colombia	507	621	30 813	0.09	0.09	5.42	1.00	1.00	0.75	0.84	1.00
Costa Rica	98	400	3 154	0.19	0.19	6.07	1.00	1.00	0.63	0.78	1.01
Croatia	589	73	456	1.42	3.63	1.10	0.96	0.96	0.91	1.20	0.99
Cyprus <sup>1</sup>	292	89	114	3.22	4.36	1.25	0.96	0.96	0.95	0.99	1.00
Dominican Republic	106	102	2 500	0.08	1.79	1.89	0.98	0.98	0.68	0.96	1.01
FYROM	19	162	451	0.12	1.67	2.84	0.98	0.98	0.95	0.96	1.00
Georgia	230	72	515	0.60	4.45	1.34	0.96	0.96	0.79	0.93	1.00
Hong Kong (China)	374	10	102	0.65	1.79	0.18	0.98	0.98	0.89	0.96	1.00
Indonesia	0	261	124 725	0.00	0.13	4.03	1.00	1.00	0.68	0.97	1.00
Jordan	1 006	448	6 256	0.92	0.97	5.70	0.99	0.99	0.86	0.92	0.98
Kazakhstan	0	0	0	0.00	3.57	0.00	0.96	0.96	0.91	0.95	1.00
Kosovo	174	215	1 010	0.77	4.84	4.49	0.95	0.95	0.71	0.84	0.99
Lebanon	0	0	0	0.00	2.09	0.00	0.98	0.98	0.66	0.70	1.00
Lithuania	1 050	68	282	3.39	5.12	0.91	0.95	0.95	0.90	0.98	1.00
Macao (China)	0	28	28	0.00	0.07	0.62	1.00	1.00	0.88	1.02	1.00
Malaysia	2 344	232	13 167	0.56	1.10	3.17	0.99	0.99	0.76	0.93	1.00
Malta	41	9	9	0.95	2.36	0.21	0.98	0.98	0.98	1.00	1.00
Moldova	118	34	194	0.40	0.99	0.66	0.99	0.99	0.93	0.98	0.99
Montenegro	332	72	78	4.66	5.17	1.10	0.95	0.95	0.90	0.97	0.98
Peru	745	329	20 685	0.17	1.50	4.78	0.99	0.99	0.74	0.92	1.00
Qatar	193	389	392	1.47	4.17	2.99	0.96	0.96	0.93	0.98	1.00
Romania	120	117	3 991	0.07	1.11	2.43	0.99	0.99	0.93	0.95	0.99
Russian Federation	2 469	32	5 732	0.22	2.28	0.51	0.98	0.98	0.95	0.94	1.04
Singapore	179	51	303	0.39	1.33	0.65	0.99	0.99	0.96	1.00	1.00
Chinese Taipei	647	80	2 420	0.26	0.67	0.96	0.99	0.99	0.85	0.88	1.00
Thailand	2 107	424	36 993	0.33	1.60	5.81	0.98	0.98	0.71	0.85	1.01
Trinidad and Tobago	0	206	421	0.00	0.00	3.19	1.00	1.00	0.76	0.76	1.00
Tunisia	61	144	2 592	0.05	0.61	2.28	0.99	0.99	0.93	0.93	1.01
United Arab Emirates	152	170	714	0.32	2.25	1.52	0.98	0.98	0.91	0.94	0.99
Uruguay	32	522	2 900	0.08	0.09	7.57	1.00	1.00	0.72	0.88	1.00
Viet Nam	0	144	24 954	0.00	0.63	2.85	0.99	0.99	0.49	0.88	0.97

1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

## [Part 1/2]

Table 11.2 PISA target populations and samples, by adjudicated regions

	All 15-year-olds	Enrolled 15-year-olds	Target population	School-level exclusions	Target minus school level exclusions	School level exclusion rate (%)	Estimation of enrolled students from frame	Number of participating students	Weighted number of participating students	Number of excluded students
<b>OECD</b>	Belgium (Flemish community)	70 451	68 173	68 173	997	67 176	1.46	65 298	5 675	62 986
	Spain (Andalusia)	88 493	82 495	82 495	251	82 244	0.30	82 193	1 813	81 642
	Spain (Aragon)	11 737	11 192	11 192	48	11 144	0.43	11 126	1 798	10 758
	Spain (Asturias)	7 391	7 186	7 186	27	7 159	0.38	7 066	1 790	6 895
	Spain (Balearic Islands)	10 629	9 623	9 623	60	9 563	0.63	9 502	1 797	9 208
	Spain (Basque Country)	18 455	18 117	18 117	60	18 057	0.33	18 113	3 612	17 424
	Spain (Canary Islands)	21 848	20 192	20 192	70	20 122	0.35	20 229	1 842	19 447
	Spain (Cantabria)	4 821	4 775	4 775	19	4 756	0.40	4 780	1 924	4 576
	Spain (Castile and Leon)	20 057	19 690	19 690	84	19 606	0.43	19 602	1 858	18 004
	Spain (CastileLaMancha)	21 165	19 646	19 646	115	19 531	0.59	19 543	1 889	19 247
	Spain (Catalonia)	70 633	68 278	68 278	612	67 666	0.90	67 606	1 769	63 112
	Spain (Extremadura)	10 955	10 745	10 745	64	10 681	0.60	10 592	1 809	10 054
	Spain (Galicia)	20 949	19 616	19 616	69	19 547	0.35	19 617	1 865	19 063
	Spain (La Rioja)	2 934	2 853	2 853	33	2 820	1.16	2 822	1 461	2 758
	Spain (Madrid)	58 569	53 865	53 865	383	53 482	0.71	53 137	1 808	53 240
	Spain (Murcia)	15 690	14 044	14 044	62	13 982	0.44	14 015	1 796	13 555
	Spain (Navarra)	6 192	5 856	5 856	27	5 829	0.46	5 793	1 874	5 496
	Spain (Valencia)	47 367	44 072	44 072	198	43 874	0.45	43 204	1 625	38 900
	United Kingdom (Scotland)	56 171	56 344	56 344	897	55 447	1.59	55 282	3 111	50 190
	United States (Massachusetts (public))	80 631	82 745	71 900	18	71 882	0.03	69 899	1 652	60 918
	United States (North Carolina (public))	130 833	116 807	110 215	416	109 799	0.38	110 786	1 887	104 161
	United States (Puerto Rico) <sup>1</sup>	50 321	44 613	44 613	760	43 853	1.70	39 453	1 398	30 261
<b>Partners</b>	Argentina (CABA)	30 974	35 767	35 767	12	35 755	0.03	35 576	1 657	32 180
	United Arab Emirates (Abu Dhabi)	19 702	19 629	19 611	204	19 407	1.04	19 402	3 610	18 335
	United Arab Emirates (Dubai)	14 662	14 643	14 642	579	14 063	3.95	14 057	6 287	12 906

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

## [Part 2/2]

Table 11.2 PISA target populations and samples, by adjudicated regions

	Weighted number of excluded students	Number of ineligible students	Weighted number of ineligible students	Within-school exclusion rate (%)	Overall exclusion rate (%)	Percentage of ineligible / withdrawn	Coverage Index 1	Coverage Index 2	Coverage Index 3	Coverage Index 4	Coverage Index 5
<b>OECD</b>	Belgium (Flemish community)	159	79	780	0.25	1.71	1.24	0.98	0.98	0.89	0.97
	Spain (Andalusia)	1 718	21	817	2.06	2.36	0.98	0.98	0.98	0.92	1.01
	Spain (Aragon)	204	20	112	1.86	2.28	1.02	0.98	0.98	0.92	0.99
	Spain (Asturias)	84	8	27	1.21	1.58	0.39	0.98	0.98	0.93	0.99
	Spain (Balearic Islands)	177	9	40	1.89	2.50	0.43	0.98	0.98	0.87	0.99
	Spain (Basque Country)	254	20	67	1.44	1.76	0.38	0.98	0.98	0.94	0.98
	Spain (Canary Islands)	374	29	285	1.89	2.23	1.44	0.98	0.98	0.89	0.98
	Spain (Cantabria)	35	8	19	0.76	1.15	0.41	0.99	0.99	0.95	0.96
	Spain (Castile and Leon)	883	14	123	4.67	5.08	0.65	0.95	0.95	0.90	0.96
	Spain (CastileLaMancha)	333	22	213	1.70	2.28	1.09	0.98	0.98	0.91	1.00
	Spain (Catalonia)	3 011	18	578	4.55	5.41	0.87	0.95	0.95	0.89	0.98
	Spain (Extremadura)	201	18	92	1.96	2.54	0.89	0.97	0.97	0.92	0.97
	Spain (Galicia)	417	3	28	2.14	2.48	0.14	0.98	0.98	0.91	0.99
	Spain (La Rioja)	7	27	48	0.26	1.41	1.73	0.99	0.99	0.94	0.98
	Spain (Madrid)	529	11	270	0.98	1.69	0.50	0.98	0.98	0.91	1.01
	Spain (Murcia)	391	4	27	2.80	3.23	0.20	0.97	0.97	0.86	1.00
	Spain (Navarra)	138	18	48	2.45	2.90	0.86	0.97	0.97	0.89	0.97
	Spain (Valencia)	3 014	12	247	7.19	7.61	0.59	0.92	0.92	0.82	0.97
	United Kingdom (Scotland)	2 645	172	2 166	5.01	6.52	4.10	0.93	0.93	0.89	1.00
	United States (Massachusetts (public))	2 785	106	3 514	4.37	4.40	5.52	0.96	0.83	0.76	0.91
	United States (North Carolina (public))	4 636	107	5 517	4.26	4.62	5.07	0.95	0.90	0.80	0.98
	United States (Puerto Rico) <sup>1</sup>	440	235	8 761	1.43	3.11	28.54	0.97	0.97	0.60	0.78
<b>Partners</b>	Argentina (CABA)	85	48	714	0.26	0.30	2.21	1.00	1.00	1.04	0.91
	United Arab Emirates (Abu Dhabi)	36	53	265	0.19	1.23	1.44	0.99	0.99	0.93	0.95
	United Arab Emirates (Dubai)	104	69	215	0.80	4.72	1.65	0.95	0.95	0.88	0.93

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

For calculating school response rates before replacement, the numerator consisted of all original sample schools with enrolled age-eligible students who participated (i.e., assessed a sample of PISA-eligible students, and obtained a student response rate of at least 50%). The denominator consisted of all the schools in the numerator, plus those original sample schools with enrolled age-eligible students that either did not participate or failed to assess at least 50% of PISA-eligible sample students. Schools that were included in the sampling frame, but were found to have no age-eligible students, or which were excluded in the field were omitted from the calculation of response rates. Replacement schools do not figure in these calculations.



Table 11.3 Response rates before school replacement

	Weighted school participation rate before replacement (%) (SCHRRW1)	Weighted number of responding schools (weighted also by enrollment) (NUMW1)	Weighted number of schools sampled (responding + non-responding) (weighted also by enrollment) (DENW1)	Unweighted school participation rate before replacement (%) (SCHRU1)	Number of responding schools (unweighted) (NUMU1)	Number of responding and non-responding schools (unweighted) (DENU1)
<b>OECD</b>						
Australia	94.42	260 657	276 072	91.37	720	788
Austria	99.95	81 690	81 730	98.53	269	273
Belgium	83.07	98 786	118 915	81.06	244	301
Canada	74.48	283 853	381 133	69.74	703	1008
Chile	92.43	215 139	232 756	89.22	207	232
Czech Republic	98.13	86 354	87 999	98.55	339	344
Denmark	90.46	57 803	63 897	88.14	327	371
Estonia	99.89	11 142	11 154	99.52	206	207
Finland	99.78	58 653	58 782	99.40	167	168
France	90.75	679 984	749 284	90.98	232	255
Germany	96.25	764 423	794 206	95.70	245	256
Greece	92.23	95 030	103 031	89.62	190	212
Hungary	93.42	83 897	89 808	92.03	231	251
Iceland	98.82	4 114	4 163	94.57	122	129
Ireland	99.29	61 023	61 461	98.82	167	169
Israel	90.90	105 192	115 717	88.95	169	190
Italy	74.39	383 933	516 113	77.82	414	532
Japan	94.45	1 087 414	1 151 305	94.50	189	200
Korea	99.65	612 937	615 107	99.41	168	169
Latvia	86.46	14 122	16 334	85.87	231	269
Luxembourg	100.00	5 891	5 891	100.00	44	44
Mexico	95.46	1 311 608	1 373 919	94.72	269	284
Netherlands	63.31	121 527	191 966	62.19	125	201
New Zealand	71.43	40 623	56 875	69.05	145	210
Norway	95.17	58 824	61 809	95.02	229	241
Poland	88.49	314 288	355 158	88.82	151	170
Portugal	85.87	87 756	102 193	83.86	213	254
Slovak Republic	92.69	50 513	54 499	92.20	272	295
Slovenia	97.69	16 886	17 286	95.13	332	349
Spain	98.87	404 640	409 246	99.00	199	201
Sweden	99.70	93 819	94 097	98.54	202	205
Switzerland	93.16	75 482	81 026	91.38	212	232
Turkey	96.88	1 057 318	1 091 317	89.74	175	195
United Kingdom	83.65	591 757	707 415	84.62	506	598
United States	66.67	2 601 386	3 902 089	66.67	142	213
<b>Partners</b>						
Albania	99.75	43 809	43 919	99.57	229	230
Algeria	96.13	341 463	355 216	95.78	159	166
Argentina	88.74	508 448	572 941	89.08	212	238
Brazil	93.19	2 509 198	2 692 686	90.66	806	889
B-S-J-G (China)	87.66	1 259 845	1 437 201	92.54	248	268
Bulgaria	99.61	56 265	56 483	99.44	179	180
Colombia	98.64	664 664	673 817	97.07	364	375
Costa Rica	99.12	66 485	67 073	99.03	204	206
Croatia	99.78	34 575	34 652	98.77	160	162
Cyprus <sup>1</sup>	96.76	8 830	9 126	92.42	122	132
Dominican Republic	98.90	136 669	138 187	98.97	193	195
FYROM	99.72	16 426	16 472	99.07	106	107
Georgia	97.49	40 552	41 595	95.88	256	267
Hong Kong (China)	75.11	45 603	60 716	75.16	115	153
Indonesia	98.44	3 126 468	3 176 076	98.31	232	236
Jordan	100.00	119 024	119 024	100.00	250	250
Kazakhstan	100.00	202 701	202 701	100.00	232	232
Kosovo	100.00	26 924	26 924	100.00	224	224
Lebanon	66.59	40 542	60 882	67.53	208	308
Lithuania	99.36	31 386	31 588	99.36	309	311
Macao (China)	100.00	4 414	4 414	100.00	45	45
Malaysia	51.39	229 340	446 237	63.91	147	230
Malta	99.95	4 341	4 343	96.72	59	61
Moldova	100.00	30 145	30 145	100.00	229	229
Montenegro	99.85	7 301	7 312	98.46	64	65
Peru	99.52	468 406	470 651	99.29	280	282
Qatar	98.98	13 333	13 470	98.81	166	168
Romania	99.36	171 553	172 652	99.45	181	182
Russia	99.37	1 181 937	1 189 441	99.52	209	210
Singapore	97.17	45 299	46 620	97.77	175	179
Chinese Taipei	100.00	286 778	286 778	100.00	214	214
Thailand	98.50	739 772	751 010	98.53	269	273
Trinidad and Tobago	91.55	15 904	17 371	86.50	141	163
Tunisia	99.17	121 751	122 767	98.18	162	165
United Arab Emirates	98.50	49 310	50 060	99.16	473	477
Uruguay	98.28	42 986	43 737	98.19	217	221
Viet Nam	100.00	996 757	996 757	100.00	188	188

1. See note 1 under Table 11.1.

Table 11.4 Response rates before school replacement, by adjudicated regions

	Weighted school participation rate before replacement (%) (SCHRRW1)	Weighted number of responding schools (weighted also by enrollment) (NUMW1)	Weighted number of schools sampled (responding + non-responding) (weighted also by enrollment) (DENW1)	Unweighted school participation rate before replacement (%) (SCHRRU1)	Number of responding schools (unweighted) (NUMU1)	Number of responding and non-responding schools (unweighted) (DENU1)	
<b>OECD</b>	Belgium (Flemish community)	75.87	49 542	65 298	74.19	138	186
	Spain (Andalusia)	98.15	80 669	82 193	98.15	53	54
	Spain (Aragon)	100.00	11 126	11 126	100.00	53	53
	Spain (Asturias)	100.00	7 066	7 066	100.00	54	54
	Spain (Balearic Islands)	100.00	9 502	9 502	100.00	54	54
	Spain (Basque Country)	100.00	18 113	18 113	100.00	119	119
	Spain (Canary Islands)	98.26	19 877	20 229	98.15	53	54
	Spain (Cantabria)	100.00	4 780	4 780	100.00	56	56
	Spain (Castile and Leon)	100.00	19 602	19 602	100.00	57	57
	Spain (CastileLaMancha)	100.00	19 543	19 543	100.00	55	55
	Spain (Catalonia)	100.00	67 606	67 606	100.00	52	52
	Spain (Extremadura)	100.00	10 592	10 592	100.00	53	53
	Spain (Galicia)	100.00	19 617	19 617	100.00	59	59
	Spain (La Rioja)	100.00	2 822	2 822	100.00	47	47
	Spain (Madrid)	97.99	52 068	53 137	98.04	50	51
	Spain (Murcia)	100.00	14 015	14 015	100.00	53	53
	Spain (Navarra)	100.00	5 793	5 793	100.00	52	52
	Spain (Valencia)	97.94	42 313	43 204	98.11	52	53
	United Kingdom (Scotland)	86.61	47 878	55 282	86.32	101	117
<b>Partners</b>	United States (Massachusetts (public))	78.40	54 800	69 899	77.36	41	53
	United States (North Carolina (public))	100.00	110 786	110 786	100.00	54	54
	United States (Puerto Rico) <sup>1</sup>	100.00	39 453	39 453	100.00	47	47

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

For calculating school response rates after replacement, the numerator consisted of all sampled schools (original plus replacement) with enrolled age-eligible students that participated (i.e., assessed a sample of PISA-eligible students and obtained a student response rate of at least 50%). The denominator consisted of all the schools in the numerator, plus those original sample schools that had age-eligible students enrolled, but that failed to assess at least 50% of PISA-eligible sample students and for which no replacement school participated. Schools that were included in the sampling frame, but were found to contain no age-eligible students, were omitted from the calculation of response rates. Replacement schools were included in rates only when they participated, and were replacing a refusing school that had age-eligible students.

In calculating weighted school response rates, each school received a weight equal to the product of its base weight (the reciprocal of its selection probability) and the number of age-eligible students enrolled in the school, as indicated on the school sampling frame.

With the use of probability proportional to size sampling, where there are no certainty or small schools, the product of the initial weight and the enrolment will be a constant, so in participating countries with few certainty school selections and no oversampling or undersampling of any explicit strata, weighted and unweighted rates are very similar. The weighted school response rate before replacement is given by the formula:

### 11.1

$$\text{weighted school response rate}_{\text{before replacement}} = \frac{\sum_{i \in Y} W_i E_i}{\sum_{i \in (Y \cup N)} W_i E_i}$$

where  $Y$  denotes the set of responding original sample schools with age-eligible students,  $N$  denotes the set of eligible non-responding original sample schools,  $W_i$  denotes the base weight for school  $i$ ,  $W_i = 1/P_i$  where  $P_i$  denotes the school selection probability for school  $i$ , and  $E_i$  denotes the enrolment size of age-eligible students, as indicated on the sampling frame.



Table 11.5 Response rates after school replacement

	Weighted school participation rate after all replacement (%) (SCHRRW3)	Weighted number of responding schools (weighted also by enrollment) (NUMW3)	Weighted number of schools sampled (responding + non-responding) (weighted also by enrollment) (DENW3)	Unweighted school participation rate after all replacement (%) (SCHRRU3)	Number of responding schools (unweighted) (NUMU3)	Number of responding and non-responding schools (unweighted) (DENU3)
<b>OECD</b>						
Australia	94.95	262 130	276 072	91.75	723	788
Austria	99.95	81 690	81 730	98.53	269	273
Belgium	95.37	113 435	118 936	95.02	286	301
Canada	78.57	299 512	381 189	72.02	726	1008
Chile	99.14	230 749	232 757	97.41	226	232
Czech Republic	98.13	86 354	87 999	98.55	339	344
Denmark	92.03	58 837	63 931	89.22	331	371
Estonia	99.89	11 142	11 154	99.52	206	207
Finland	100.00	58 800	58 800	100.00	168	168
France	94.34	706 838	749 284	94.51	241	255
Germany	98.94	785 813	794 206	98.83	253	256
Greece	98.48	101 653	103 218	98.58	209	212
Hungary	98.80	88 751	89 825	97.21	244	251
Iceland	98.82	4 114	4 163	94.57	122	129
Ireland	99.29	61 023	61 461	98.82	167	169
Israel	92.96	107 570	115 717	91.05	173	190
Italy	87.50	451 098	515 515	87.22	464	532
Japan	98.99	1 139 734	1 151 305	99.00	198	200
Korea	99.65	612 937	615 107	99.41	168	169
Latvia	92.52	15 103	16 324	92.19	248	269
Luxembourg	100.00	5 891	5 891	100.00	44	44
Mexico	97.52	1 339 901	1 373 919	96.83	275	284
Netherlands	93.21	178 929	191 966	91.54	184	201
New Zealand	84.50	48 094	56 913	83.81	176	210
Norway	95.17	58 824	61 809	95.02	229	241
Poland	99.32	352 754	355 158	98.82	168	170
Portugal	95.10	97 516	102 537	93.70	238	254
Slovak Republic	98.80	53 908	54 562	97.63	288	295
Slovenia	97.75	16 896	17 286	95.42	333	349
Spain	100.00	409 246	409 246	100.00	201	201
Sweden	99.70	93 819	94 097	98.54	202	205
Switzerland	97.67	79 481	81 375	96.98	225	232
Turkey	99.12	1 081 935	1 091 528	95.90	187	195
United Kingdom	92.59	654 992	707 415	91.47	547	598
United States	83.32	3 244 399	3 893 828	83.10	177	213
<b>Partners</b>						
Albania	99.75	43 809	43 919	99.57	229	230
Algeria	96.13	341 463	355 216	95.78	159	166
Argentina	97.13	556 478	572 941	97.06	231	238
Brazil	94.08	2 533 711	2 693 137	91.68	815	889
B-S-J-G (China)	100.00	1 437 652	1 437 652	100.00	268	268
Bulgaria	100.00	56 600	56 600	100.00	180	180
Colombia	99.81	672 526	673 835	98.93	371	375
Costa Rica	99.12	66 485	67 073	99.03	204	206
Croatia	99.78	34 575	34 652	98.77	160	162
Cyprus <sup>1</sup>	96.76	8 830	9 126	92.42	122	132
Dominican Republic	98.90	136 669	138 187	98.97	193	195
FYROM	99.72	16 426	16 472	99.07	106	107
Georgia	98.83	41 081	41 566	98.13	262	267
Hong Kong (China)	90.25	54 795	60 715	90.20	138	153
Indonesia	100.00	3 176 076	3 176 076	100.00	236	236
Jordan	100.00	119 024	119 024	100.00	250	250
Kazakhstan	100.00	202 701	202 701	100.00	232	232
Kosovo	100.00	26 924	26 924	100.00	224	224
Lebanon	87.33	53 091	60 797	87.66	270	308
Lithuania	99.86	31 543	31 588	99.68	310	311
Macao (China)	100.00	4 414	4 414	100.00	45	45
Malaysia	98.06	437 424	446 100	97.39	224	230
Malta	99.95	4 341	4 343	96.72	59	61
Moldova	100.00	30 145	30 145	100.00	229	229
Montenegro	99.85	7 301	7 312	98.46	64	65
Peru	99.79	469 662	470 651	99.65	281	282
Qatar	98.98	13 333	13 470	98.81	166	168
Romania	100.00	172 495	172 495	100.00	182	182
Russia	99.37	1 181 937	1 189 441	99.52	209	210
Singapore	97.71	45 553	46 620	98.32	176	179
Chinese Taipei	100.00	286 778	286 778	100.00	214	214
Thailand	100.00	751 010	751 010	100.00	273	273
Trinidad and Tobago	91.55	15 904	17 371	86.50	141	163
Tunisia	99.22	121 838	122 792	98.79	163	165
United Arab Emirates	98.50	49 310	50 060	99.16	473	477
Uruguay	99.33	43 442	43 737	99.10	219	221
Viet Nam	100.00	996 757	996 757	100.00	188	188

1. See note 1 under Table 11.1.

Table 11.6 Response rates after school replacement, by adjudicated regions

	Weighted school participation rate after all replacement (%) (SCHRRW3)	Weighted number of responding schools (weighted also by enrollment) (NUMW3)	Weighted number of schools sampled (responding + non-responding) (weighted also by enrollment) (DENW3)	Unweighted school participation rate after all replacement (%) (SCHRU3)	Number of responding schools (unweighted) (NUMU3)	Number of responding and non-responding schools (unweighted) (DENU3)	
<b>OECD</b>	Belgium (Flemish community)	93.45	61 039.32	65 319.22	93.55	174	186
	Spain (Andalusia)	100.00	82 192.73	82 192.73	100.00	54	54
	Spain (Aragon)	100.00	11 125.90	11 125.90	100.00	53	53
	Spain (Asturias)	100.00	7 065.53	7 065.53	100.00	54	54
	Spain (Balearic Islands)	100.00	9 501.65	9 501.65	100.00	54	54
	Spain (Basque Country)	100.00	18 113.27	18 113.27	100.00	119	119
	Spain (Canary Islands)	98.26	19 877.44	20 229.40	98.15	53	54
	Spain (Cantabria)	100.00	4 779.92	4 779.92	100.00	56	56
	Spain (Castile and Leon)	100.00	19 601.83	19 601.83	100.00	57	57
	Spain (CastileLaMancha)	100.00	19 542.72	19 542.72	100.00	55	55
	Spain (Catalonia)	100.00	67 606.13	67 606.13	100.00	52	52
	Spain (Extremadura)	100.00	10 592.13	10 592.13	100.00	53	53
	Spain (Galicia)	100.00	19 616.86	19 616.86	100.00	59	59
	Spain (La Rioja)	100.00	2 822.00	2 822.00	100.00	47	47
	Spain (Madrid)	100.00	53 137.04	53 137.04	100.00	51	51
	Spain (Murcia)	100.00	14 015.27	14 015.27	100.00	53	53
	Spain (Navarra)	100.00	5 793.20	5 793.20	100.00	52	52
	Spain (Valencia)	97.94	42 313.15	43 203.77	98.11	52	53
<b>Partners</b>	United Kingdom (Scotland)	92.68	51 235.75	55 282.20	92.31	108	117
	United States (Massachusetts (public))	91.85	64 205.61	69 899.08	90.57	48	53
	United States (North Carolina (public))	100.00	110 785.88	110 785.88	100.00	54	54
	United States (Puerto Rico) <sup>1</sup>	100.00	39 453.16	39 453.16	100.00	47	47

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

The weighted school response rate, after replacement, is given by the formula:

### 11.2

$$\text{weighted school response rate}_{\text{after replacement}} = \frac{\sum_{i \in (YUR)} W_i E_i}{\sum_{i \in (YURUN)} W_i E_i}$$

where  $Y$  denotes the set of responding original sample schools,  $R$  denotes the set of responding replacement schools, for which the corresponding original sample school was eligible but was non-responding,  $N$  denotes the set of eligible refusing original sample schools,  $W_i$  denotes the base weight for school  $i$ ,  $W_i = 1/P_i$ , where  $P_i$  denotes the school selection probability for school  $i$ , and for weighted rates,  $E_i$  denotes the enrolment size of age-eligible students, as indicated on the sampling frame.

For unweighted student response rates, the numerator is the number of students for whom assessment data were included in the results less those in schools with between 25 and 50% student participation. The denominator is the number of sampled students who were age-eligible, and not explicitly excluded as student exclusions.

For weighted student response rates, the same number of students appears in the numerator and denominator as for unweighted rates, but each student was weighted by its student base weight. This is given as the product of the school base weight – for the school in which the student was enrolled – and the reciprocal of the student selection probability within the school.

In countries with no oversampling of any explicit strata, weighted and unweighted student participation rates are very similar.

Overall response rates are calculated as the product of school and student response rates. Although overall weighted and unweighted rates can be calculated, there is little value in presenting overall unweighted rates. The weighted rates indicate the proportion of the student population represented by the sample prior to making the school and student non-response adjustments.



Table 11.7 Response rates, students within schools after school replacement

	Weighted student participation rate after second replacement (%) (STURRW3)	Number of students assessed (Weighted) (NUMSTW3)	Number of students sampled (assessed + absent) (weighted) (DENSTW3)	Unweighted student participation rate after second replacement (%) (STURRU3)	Number of students assessed (unweighted) (NUMSTU3)	Number of students sampled (assessed + absent) (unweighted) (DENSTU3)
<b>OECD</b>						
Australia	83.99	204 763	243 789	80.61	14 089	17 477
Austria	86.59	63 660	73 521	71.01	7 007	9 868
Belgium	90.63	99 760	110 075	90.88	9 635	10 602
Canada	80.80	210 476	260 487	81.25	19 604	24 129
Chile	93.31	189 206	202 774	93.67	7 039	7 515
Czech Republic	88.77	73 386	82 672	88.85	6 835	7 693
Denmark	89.08	49 732	55 830	87.35	7 149	8 184
Estonia	93.22	10 088	10 822	93.21	5 587	5 994
Finland	93.44	53 198	56 934	93.45	5 882	6 294
France	88.21	611 563	693 336	88.16	5 980	6 783
Germany	93.27	685 972	735 487	93.26	6 476	6 944
Greece	94.32	89 588	94 986	94.40	5 511	5 838
Hungary	92.30	77 212	83 657	92.49	5 643	6 101
Iceland	86.11	3 365	3 908	86.11	3 365	3 908
Ireland	88.60	51 947	58 630	88.62	5 741	6 478
Israel	90.48	98 572	108 940	90.46	6 598	7 294
Italy	87.67	377 011	430 041	89.38	11 477	12 841
Japan	97.24	1 096 193	1 127 265	97.21	6 647	6 838
Korea	98.56	559 121	567 284	98.53	5 581	5 664
Latvia	90.42	12 799	14 155	90.26	4 845	5 368
Luxembourg	95.65	5 299	5 540	95.65	5 299	5 540
Mexico	95.43	1 290 435	1 352 237	95.34	7 568	7 938
Netherlands	85.12	152 346	178 985	85.26	5 345	6 269
New Zealand	80.31	36 860	45 897	80.28	4 453	5 547
Norway	90.75	50 163	55 277	90.69	5 456	6 016
Poland	87.54	300 617	343 405	87.43	4 466	5 108
Portugal	82.02	75 391	91 916	82.23	7 180	8 732
Slovak Republic	92.37	45 357	49 103	91.91	6 342	6 900
Slovenia	91.77	15 072	16 424	91.40	6 406	7 009
Spain	89.14	356 509	399 935	89.34	6 736	7 540
Sweden	90.67	82 582	91 081	90.77	5 458	6 013
Switzerland	92.45	74 465	80 544	92.59	5 838	6 305
Turkey	95.19	874 609	918 816	94.91	5 895	6 211
United Kingdom	89.02	517 426	581 252	87.58	14 120	16 123
United States	89.76	2 629 707	2 929 771	89.59	5 712	6 376
<b>Partners</b>						
Albania	93.53	38 174	40 814	93.84	5 213	5 555
Algeria	92.47	274 121	296 434	92.59	5 494	5 934
Argentina	90.36	345 508	382 352	89.95	6 311	7 016
Brazil	87.32	1 996 574	2 286 505	85.73	22 791	26 586
B-S-J-G (China)	96.69	1 287 710	1 331 794	97.46	9 841	10 097
Bulgaria	94.87	50 931	53 685	95.00	5 928	6 240
Colombia	94.52	535 682	566 734	93.39	11 777	12 611
Costa Rica	92.46	47 494	51 369	92.38	6 846	7 411
Croatia	91.35	37 275	40 803	91.42	5 809	6 354
Cyprus*	94.03	8 016	8 526	93.35	5 561	5 957
Dominican Republic	93.82	122 620	130 700	94.13	4 731	5 026
FYROM	94.92	14 999	15 802	94.78	5 324	5 617
Georgia	93.91	35 567	37 873	93.44	5 316	5 689
Hong Kong (China)	93.08	48 222	51 806	93.25	5 359	5 747
Indonesia	97.51	3 015 844	3 092 773	97.30	6 513	6 694
Jordan	97.42	105 868	108 669	97.39	7 267	7 462
Kazakhstan	97.29	187 683	192 921	97.29	7 841	8 059
Kosovo	98.58	22 016	22 333	98.57	4 826	4 896
Lebanon	94.52	36 052	38 143	94.95	4 546	4 788
Lithuania	90.57	27 070	29 889	90.57	6 523	7 202
Macao (China)	99.31	4 476	4 507	99.31	4 476	4 507
Malaysia	96.66	393 785	407 396	97.21	8 843	9 097
Malta	84.63	3 634	4 294	84.63	3 634	4 294
Moldova	98.00	28 754	29 341	97.96	5 325	5 436
Montenegro	93.79	6 346	6 766	93.74	5 665	6 043
Peru	98.90	426 205	430 959	98.82	6 971	7 054
Qatar	94.09	12 061	12 819	94.09	12 061	12 819
Romania	99.21	162 918	164 216	99.31	4 876	4 910
Russia	96.83	1 072 914	1 108 068	96.88	6 021	6 215
Singapore	93.33	42 241	45 259	93.14	6 105	6 555
Chinese Taipei	98.00	246 408	251 424	97.93	7 708	7 871
Thailand	96.88	614 996	634 795	97.15	8 249	8 491
Trinidad and Tobago	79.38	9 674	12 188	79.84	4 587	5 745
Tunisia	86.40	97 337	112 665	86.48	5 340	6 175
United Arab Emirates	94.62	43 774	46 263	94.36	14 167	15 014
Uruguay	86.16	32 762	38 023	86.24	6 059	7 026
Viet Nam	99.60	871 353	874 859	99.61	5 826	5 849

\* See note 1 under Table 11.1.

Table 11.8 Response rates, students within schools after school replacement, by adjudicated regions

	Weighted student participation rate after second replacement (%) (STURRW3)	Number of students assessed (weighted) (NUMSTW3)	Number of students sampled (assessed + absent) (weighted) (DENSTW3)	Unweighted student participation rate after second replacement (%) (STURRU3)	Number of students assessed (Unweighted) (NUMSTU3)	Number of students sampled (assessed + absent) (unweighted) (DENSTU3)
<b>OECD</b>	Belgium (Flemish community)	91.54	54 082.90	59 081.47	5 674	6 199
	Spain (Andalusia)	87.64	71 549.56	81 642.36	1 813	2 065
	Spain (Aragon)	89.49	9 626.75	10 757.56	1 798	2 008
	Spain (Asturias)	89.63	6 179.65	6 894.55	1 790	1 995
	Spain (Balearic Islands)	88.84	8 179.56	9 207.58	1 797	2 021
	Spain (Basque Country)	91.07	15 868.19	17 424.20	3 612	3 992
	Spain (Canary Islands)	90.40	17 279.43	19 113.67	1 825	2 019
	Spain (Cantabria)	90.39	4 136.09	4 575.66	1 924	2 124
	Spain (Castile and Leon)	92.03	16 568.49	18 003.77	1 858	2 020
	Spain (CastileLaMancha)	90.24	17 368.92	19 247.29	1 889	2 092
	Spain (Catalonia)	90.66	57 218.40	63 112.16	1 769	1 950
	Spain (Extremadura)	89.90	9 038.97	10 054.22	1 809	2 012
	Spain (Galicia)	91.13	17 371.25	19 062.58	1 865	2 048
	Spain (La Rioja)	91.71	2 529.21	2 757.90	1 461	1 590
	Spain (Madrid)	89.77	47 792.04	53 239.55	1 808	2 009
	Spain (Murcia)	86.96	11 787.15	13 555.12	1 796	2 064
	Spain (Navarra)	94.02	5 166.61	5 495.51	1 874	1 990
	Spain (Valencia)	87.50	33 270.94	38 024.57	1 611	1 840
<b>Partners</b>	United Kingdom (Scotland)	79.99	37 114.07	46 396.20	3 095	3 869
	United States (Massachusetts (public))	90.36	42 557.08	47 096.94	1 391	1 534
	United States (North Carolina (public))	92.43	96 277.78	104 161.17	1 887	2 038
	United States (Puerto Rico) <sup>1</sup>	93.12	28 179.19	30 261.01	1 398	1 493
	Argentina (CABA)	90.34	28 282.38	31 306.97	1 649	1 846
	United Arab Emirates (Abu Dhabi)	93.40	16 483.27	17 647.64	3 610	3 878
	United Arab Emirates (Dubai)	94.34	12 174.95	12 905.86	6 287	6 677

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

## TEACHER RESPONSE RATES

Unweighted response rates for both science and non-science teachers were created using similar methods to those for unweighted student and school response rates – that is, ineligible teachers are not used in the denominator for the rate calculation.

These rates are presented in Table 11.9 for science teachers and in Table 11.10 for the non-science teachers.

In addition to these rates, unweighted response rates were calculated also for each sampled school in each country which implemented the Teacher Questionnaire. These rates were created as quality indicators for the questionnaire team who would use the Teacher Questionnaire data to create derived variables to help provide context about PISA students.

Table 11.9 Science teacher response rates

	Country	Science teacher unweighted response rate (%)	Science teacher numerator	Science teacher denominator	Number of ineligible science teachers
<b>OECD</b>	Australia	73.49	4 158	5 658	72
	Chile	90.07	771	856	110
	Czech Republic	94.88	2 169	2 286	18
	Germany	68.90	2 032	2 949	0
	Italy	74.50	2 422	3 251	23
	Korea	99.36	926	932	4
	Portugal	91.20	1 441	1 580	29
	Spain	95.53	1 368	1 432	33
	United States	87.20	1 110	1 273	12
	United States (Massachusetts (public))	90.49	390	431	9
<b>Partners</b>	United States (North Carolina (public))	97.19	380	391	2
	Brazil	70.35	2 650	3 767	0
	B-S-J-G (China)	99.30	2 410	2 427	29
	Colombia	85.42	1 324	1 550	57
	Dominican Republic	91.13	452	496	33
	Hong Kong (China)	91.48	1 042	1 139	4
	Macao (China)	98.99	391	395	2
	Malaysia	97.67	2 010	2 058	41
	Peru	95.65	902	943	33
	Chinese Taipei	98.98	1 545	1 561	9
	United Arab Emirates	89.13	1 795	2 014	10
	United Arab Emirates (Abu Dhabi)	87.83	729	830	7
	United Arab Emirates (Dubai)	90.34	1 103	1 221	7



Table 11.10 Non-science teacher response rates

Country	Non-Science teacher unweighted response rate (%)	Non-Science teacher numerator	Non-Science teacher denominator	Number of ineligible non-science teachers
<b>OECD</b>	Australia	71.25	7 394	10 378
	Chile	90.68	2 295	2 531
	Czech Republic	93.75	3 750	4 000
	Germany	64.90	3 568	5 498
	Italy	70.45	4 526	6 424
	Korea	99.12	2 128	2 147
	Portugal	88.20	2 257	2 559
	Spain	92.46	2 526	2 732
	United States	88.53	2 099	2 371
United States (Massachusetts (public))	89.36	630	705	10
	United States (North Carolina (public))	95.47	738	773
<b>Partners</b>	Brazil	67.01	5 398	8 055
	B-S-J-G (China)	99.03	3 880	3 918
	Colombia	82.89	3 295	3 975
	Dominican Republic	86.97	1 048	1 205
	Hong Kong (China)	89.80	1 841	2 050
	Macao (China)	99.34	2 410	2 426
	Malaysia	97.44	3 191	3 275
	Peru	99.32	2 918	2 938
	Chinese Taipei	99.08	3 130	3 159
	United Arab Emirates	87.23	3 285	3 766
	United Arab Emirates (Abu Dhabi)	87.29	1 222	1 400
	United Arab Emirates (Dubai)	88.78	2 026	2 282

## DESIGN EFFECTS AND EFFECTIVE SAMPLE SIZES

Surveys in education and especially international surveys rarely sample students by simply selecting a random sample of students (known as a simple random sample, or SRS). Rather, a sampling design is used where schools are first selected and, within each selected school, classes or students are randomly sampled. Sometimes, geographic areas are first selected before sampling schools and students. This sampling design is usually referred to as a cluster sample or a multi-stage sample.

Selected students attending the same school cannot be considered as independent observations as assumed with a simple random sample because they are usually more similar to one another than to students attending other schools. For instance, the students are offered the same school resources, may have the same teachers and therefore are taught a common implemented curriculum, and so on. School differences are also larger if different educational programmes are not available in all schools. One expects to observe greater differences between a vocational school and an academic school than between two comprehensive schools.

Furthermore, it is well known that within a country, within sub-national entities and within a city, people tend to live in areas according to their financial resources. As children usually attend schools close to their home, it is likely that students attending the same school come from similar social and economic backgrounds.

A simple random sample of 4 000 students is thus likely to cover the diversity of the population better than a sample of 100 schools with 40 students observed within each school. It follows that the uncertainty associated with any population parameter estimate (i.e., standard error) will be larger for a clustered sample estimate than for a simple random sample estimate of the same size.

In the case of a simple random sample, the standard error of a mean estimate is equal to:

**11.3**

$$\sigma_{(\bar{\mu})} = \sqrt{\frac{\sigma^2}{n}}$$

where  $\sigma^2$  denotes the variance of the whole student population and  $n$  is the student sample size.



For an infinite population of schools and infinite populations of students within schools, the standard error of a mean estimate from a cluster sample is equal to:

#### 11.4

$$\sigma_{(\bar{\mu})} = \sqrt{\frac{\sigma_{schools}^2}{n_{schools}} + \frac{\sigma_{within}^2}{n_{schools} n_{students}}}$$

where  $\sigma_{schools}^2$  denotes the variance of the school means,  $\sigma_{within}^2$  denotes the variances of students within schools,  $n_{schools}$  denotes the sample size of schools, and  $n_{students}$  denotes the sample size of students within each school.

The standard error for the mean from a simple random sample is inversely proportional to the square root of the number of selected students. The standard error for the mean from a cluster sample is proportional to the variance that lies between clusters (i.e. schools) and within clusters and inversely proportional to the square root of the number of selected schools and is also a function of the number of students selected per school.

It is usual to express the decomposition of the total variance into the between-school variance and the within-school variance by the coefficient of intraclass correlation, also denoted *Rho*. Mathematically, this index is equal to:

#### 11.5

$$Rho = \frac{\sigma_{schools}^2}{\sigma_{schools}^2 + \sigma_{within}^2}$$

This index provides an indication of the percentage of variance that lies between schools. A low intraclass correlation indicates that schools are performing similarly while higher values point towards large differences between school performance.

To limit the reduction of precision in the population parameter estimate, multi-stage sample designs usually use supplementary information to improve coverage of the population diversity. In PISA the following techniques were implemented to limit the increase in the standard error: (i) explicit and implicit stratification of the school sampling frame and (ii) selection of schools with probabilities proportional to their size. Complementary information generally cannot compensate totally for the increase in the standard error due to the multi-stage design however but will greatly reduce it.

Table 11.11 provides the standard errors on the PISA 2015 main domain scales, calculated as if the participating country sample was selected according to (i) a simple random sample; (ii) a multi-stage procedure without using complementary information (unstratified multi-stage sampling, with sampling weights ignored) and (iii) the unbiased BRR estimate for the actual PISA 2015 design, using Fay's method. It should be mentioned that the plausible value imputation variance was not included in these computations, which thus only reflect sampling error.

Note that the values in Table 11.11 for the standard errors for the unstratified multi-stage design are overestimates for countries that had a school census (Cyprus<sup>1</sup>, Iceland, Luxembourg, Macao (China), Malta, Trinidad and Tobago, and Qatar) since these standard error estimates assume a sample of schools was collected.

Also note that in some of the countries where the BRR estimates in Table 11.11 are greater than the values for the unstratified multi-stage sample, this is because of regional or other oversampling (The countries with oversampling were: Argentina, Australia, Belgium, Brazil, B-S-J-G (China), Canada, Colombia, the Czech Republic, Denmark, Italy, Malaysia, Portugal, the United Arab Emirates, the United Kingdom).

The BRR estimates in Table 11.11 are also greater than the values for the unstratified multi-stage sample for almost all countries since nearly every country undersamples very small schools. As described in the sampling design chapter, some countries have a substantial proportion of students attending schools that have fewer students than the target cluster size (TCS). When small school undersampling was done, very small schools were undersampled while all other schools were slightly oversampled in compensation. In such cases, very small schools with at most 0, 1, or 2 age-eligible PISA students expected to be enrolled were typically undersampled by a factor of 4 while the very small schools with between 3 and TCS/2 age-eligible PISA students expected to be enrolled were undersampled by a factor of 2. This takes the allocation of schools to strata slightly away from proportional allocation, which can add slightly to weight variability and therefore to sampling variance. This is done though, to help countries minimise the operational burden of having too many small schools in their sample.

For the other instances of countries in Table 11.11 that have BRR estimates that are somewhat greater than estimates based on an unstratified multi-stage design it is unclear why the BRR variance should be larger, though it is possible that the stratification undertaken possibly did not explain enough between-school variance in these countries.

1. See note 1 under Table 11.1.



It is usual to express the effect of the sampling design on the standard errors by a statistic referred to as the design effect. This corresponds to the ratio of the variance of the estimate obtained from the (more complex) sample to the variance of the estimate that would be obtained from a simple random sample of the same number of sampling units. The design effect has two primary uses – in sample size estimation and in appraising the efficiency of more complex sampling plans (Cochran, 1977).

In PISA, as sampling variance has to be estimated by using the 80 *BRR* replicates, a design effect can be computed for a statistic  $t$  using:

**11.6**

$$Deff(t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)}$$

where  $Var_{BRR}(t)$  is the sampling variance for the statistic  $t$  computed by the *BRR* replication method, and  $Var_{SRS}(t)$  is the sampling variance for the same statistic  $t$  on the same data but considering the sample as a simple random sample.

Based on Table 11.11, the unbiased *BRR* standard error on the mean estimate in science in Australia (for example) is equal to 1.46 (rounded from 1.45568). As the standard deviation of the science performance is equal to 102.29735, the design effect in Australia for the mean estimate in science is therefore equal to:

**11.7**

$$Deff(t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)} = \frac{(1.45568)^2}{[102.29735^2 / 14\ 530]} = 2.942195$$

The sampling variance on the science performance mean in Australia is about 2.94 times larger than it would have been with a simple random sample of the same sample size. Note that the participating students are 14 530 as this number were assessed for science.

Another way to express the reduction of precision due to the complex sampling design is through the effective sample size, which expresses the simple random sample size that would give the same sampling variance as the one obtained from the actual complex sample design. The effective sample size for a statistic  $t$  is equal to:

**11.8**

$$Effn(t) = \frac{n}{Deff(t)} = \frac{n \times Var_{SRS}(t)}{Var_{BRR}(t)}$$

where  $n$  is equal to the actual number of units in the sample. The effective sample size in Australia for the science performance mean is equal to:

**11.9**

$$Effn(t) = \frac{n}{Deff(t)} = \frac{14\ 530}{2.942195} = 4938.4898$$

In other words, a simple random sample of 4 938 students in Australia would have been as precise as the actual PISA 2015 sample for the national estimate of mean science proficiency.

## VARIABILITY OF THE DESIGN EFFECT

Neither the design effect nor the effective sample size is a definitive characteristic of a sample. Both the design effect and the effective sample size vary with the variable and statistic of interest.

As previously stated, the sampling variance for estimates of the mean from a cluster sample is proportional to the intraclass correlation. In some countries, student performance varies between schools. Students in academic schools usually tend to perform well while on average student performance in vocational schools is lower. Let us now suppose that the height of the students was also measured, and there are no reasons why students in academic schools should be of different height than students in vocational schools. For this particular variable, the expected value of the between-school variance should be equal to zero and therefore, the design effect should tend to one. As the segregation effect differs according to the variable, the design effect will also differ according to the variable.

The second factor that influences the size of the design effect is the choice of requested statistics. It tends to be large for means, proportions, and sums but substantially smaller for bivariate or multivariate statistics such as correlation and regression coefficients.

## Design effects in PISA for performance variables

The notion of design effect as given earlier is extended and gives rise to five different design effect formulae to describe the influence of the sampling and test designs on the standard errors for statistics.

The total errors computed for the international PISA initial reports (OECD, 2016a,b) that involves performance variables (scale scores) consist of two components: sampling variance and measurement variance. The standard error of proficiency estimates in PISA is inflated because the students were not sampled according to a simple random sample and also because the estimation of student proficiency includes some amount of measurement error.

For any statistic  $r$ , the population estimate and the sampling variance are computed for each plausible value and then combined as described in Chapter 9.

The five design effects and their respective effective sample sizes are defined as follows:

- Design Effect 1

**11.10**

$$Deff_1(r) = \frac{Var_{SRS}(r) + MVar(r)}{Var_{SRS}(r)}$$

where  $MVar(r)$  is the measurement variance for the statistic  $r$ . This design effect shows the inflation of the total variance that would have occurred due to measurement error if in fact the samples were considered as a simple random sample.

- Design Effect 2

**11.11**

$$Deff_2(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{SRS}(r) + MVar(r)}$$

shows the inflation of the *total* variance due only to the use of a complex sampling design.

- Design Effect 3

**11.12**

$$Deff_3(r) = \frac{Var_{BRR}(r)}{Var_{SRS}(r)}$$

shows the inflation of the sampling variance due to the use of a complex design.

- Design Effect 4

**11.13**

$$Deff_4(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{BRR}(r)}$$

shows the inflation of the total variance due to measurement variance.

- Design Effect 5

**11.14**

$$Deff_5(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{SRS}(r)}$$

shows the inflation of the total variance due to the measurement variance and due to the complex sampling design.

The product of the first and second design effects equals the product of the third and fourth design effects, and both products are equal to the fifth design effect.

Tables 11.12 through 11.16 present the values of the different design effects and the corresponding effective sample sizes for each of the major domains.



Table 11.11 Standard errors for the PISA 2015 main domain scales

Country	Collaborative problem solving		Financial literacy		Mathematical literacy		Reading literacy		Science literacy	
	Simple random sample	Unbiased BRR	Simple random sample	Unbiased BRR	Simple random sample	Unbiased BRR	Simple random sample	Unbiased BRR	Simple random sample	Unbiased BRR
<b>OECD</b>										
Australia	0.88	1.52	0.98	1.84	0.77	1.33	0.85	1.36	0.85	1.46
Austria	1.18	2.34			1.14	2.68	1.21	2.57	1.16	2.40
Belgium	1.00	2.24	1.41	2.61	0.99	2.27	1.02	2.34	1.02	2.27
Canada	0.74	2.08	0.93	3.65	0.62	2.14	0.66	2.15	0.65	2.06
Chile	1.00	2.28	1.20	2.97	1.02	2.36	1.05	2.51	1.02	2.33
Czech Republic	1.10	1.99			1.09	2.23	1.21	2.48	1.15	2.25
Denmark	1.07	2.34			0.95	2.01	1.03	2.41	1.07	2.35
Estonia	1.21	2.02			1.08	1.78	1.17	2.01	1.19	1.96
Finland	1.32	2.30			1.07	2.03	1.22	2.51	1.25	2.36
France	1.28	1.93			1.22	1.98	1.43	2.36	1.30	2.03
Germany	1.25	2.48			1.10	2.45	1.24	2.89	1.23	2.63
Greece	1.24	3.47			1.20	3.56	1.32	4.27	1.24	3.89
Hungary	1.27	2.25			1.25	2.35	1.29	2.49	1.28	2.38
Iceland	1.63	1.72			1.60	1.68	1.71	1.80	1.57	1.66
Ireland					1.05	2.00	1.14	2.27	1.17	2.29
Israel	1.36	3.52			1.27	3.41	1.39	3.73	1.31	3.42
Italy	0.89	2.42	0.85	2.42	0.87	2.63	0.87	2.43	0.85	2.46
Japan	1.04	2.55			1.08	2.77	1.13	3.11	1.15	2.94
Korea	1.12	2.23			1.33	3.49	1.30	3.25	1.27	3.09
Latvia	1.29	1.74			1.11	1.54	1.21	1.64	1.18	1.46
Luxembourg	1.37	1.07			1.29	0.82	1.46	0.96	1.38	0.86
Mexico	0.91	2.21			0.86	2.21	0.90	2.37	0.82	2.06
Netherlands	1.32	2.24	1.53	2.51	1.25	2.08	1.38	2.22	1.38	2.22
New Zealand	1.57	2.19			1.37	2.11	1.56	2.26	1.55	2.35
Norway	1.27	2.22			1.15	2.05	1.34	2.34	1.30	2.23
Poland			1.48	2.70	1.31	2.31	1.34	2.26	1.36	2.48
Portugal	1.07	2.38			1.12	2.41	1.07	2.47	1.07	2.35
Puerto Rico (United States) <sup>1</sup>					2.06	5.35	2.56	6.94	2.31	6.00
Slovak Republic	1.17	2.27	1.44	3.38	1.20	2.47	1.31	2.71	1.24	2.56
Slovenia	1.16	1.34			1.10	1.14	1.15	1.16	1.19	1.23
Spain	1.07	1.96	1.21	2.74	1.03	2.02	1.06	2.18	1.07	2.05
Sweden	1.33	3.22			1.22	3.06	1.38	3.40	1.39	3.53
Switzerland					1.25	2.80	1.28	2.89	1.30	2.86
Turkey	1.02	3.38			1.07	4.08	1.07	3.91	1.03	3.88
United Kingdom	0.87	2.47			0.78	2.42	0.81	2.51	0.84	2.47
United States	1.43	3.44	1.35	3.49	1.17	3.07	1.32	3.32	1.30	3.13
<b>Partners</b>										
Albania					1.19	3.37	1.34	4.00	1.09	3.20
Algeria					0.96	2.83	0.98	2.84	0.93	2.56
Argentina					1.01	3.00	1.11	3.17	1.01	2.75
Brazil	0.58	2.11	0.72	3.17	0.59	2.55	0.66	2.44	0.59	2.27
B-S-J-G (China)	0.98	3.90	1.18	5.40	1.07	4.74	1.10	5.08	1.04	4.62
Bulgaria	1.27	3.79			1.26	3.88	1.49	4.87	1.32	4.34
Colombia	0.76	2.27			0.71	2.15	0.83	2.79	0.74	2.31
Costa Rica	0.94	2.17			0.83	2.12	0.96	2.57	0.85	2.04
Croatia	1.14	2.36			1.16	2.56	1.19	2.59	1.17	2.42
Cyprus <sup>2</sup>	1.22	1.25			1.24	1.13	1.37	1.32	1.24	1.22
Dominican Republic					1.00	2.29	1.23	2.94	1.05	2.45
FYROM					1.31	1.16	1.36	1.17	1.16	1.08
Georgia					1.29	2.61	1.42	2.76	1.24	2.36
Hong Kong (China)	1.24	2.75			1.23	2.87	1.17	2.59	1.10	2.43
Indonesia					0.99	2.91	0.94	2.72	0.85	2.49
Jordan					1.01	2.45	1.10	2.71	0.99	2.62
Kazakhstan					0.93	3.90	0.91	3.11	0.86	3.61
Kosovo					1.08	1.47	1.13	1.42	1.03	1.37
Lebanon					1.50	3.57	1.71	4.22	1.34	3.31
Lithuania	1.12	2.31	1.19	2.77	1.07	2.23	1.17	2.68	1.13	2.57
Macao (China)	1.34	0.96			1.19	0.89	1.23	0.87	1.22	0.90
Malaysia	0.85	3.22			0.85	3.11	0.86	3.37	0.80	2.95
Malta					1.83	1.43	2.00	1.54	1.95	1.45
Moldova					1.24	2.25	1.34	2.41	1.18	1.90
Montenegro	1.05	0.94			1.15	1.02	1.25	1.10	1.13	0.98
Peru	1.00	2.38	1.23	3.07	0.99	2.43	1.07	2.76	0.92	2.30
Qatar					0.90	0.67	1.01	0.77	0.90	0.71
Romania					1.24	3.70	1.36	3.99	1.13	3.21
Russia	1.19	3.28	1.11	3.07	1.07	2.99	1.13	2.94	1.06	2.90
Singapore	1.24	1.07			1.22	1.15	1.26	1.23	1.32	1.11
Chinese Taipei	1.03	2.29			1.17	2.68	1.06	2.42	1.13	2.62
Thailand	0.92	3.35			0.90	2.94	0.88	3.21	0.86	2.79
Trinidad and Tobago					1.40	1.05	1.52	1.24	1.37	1.12
Tunisia	0.80	1.84			1.15	2.84	1.11	2.61	0.88	2.01
United Arab Emirates	0.80	2.28			0.81	2.20	0.89	2.67	0.83	2.40
Uruguay	1.17	2.17			1.11	2.16	1.24	2.42	1.11	2.17
Viet Nam					1.10	4.38	0.95	3.67	1.00	3.86

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

2. See note 1 under Table 11.1.

Table 11.12 Design effects and effective sample sizes for scientific literacy

Country	Design effect 1	Design effect 2	Design effect 3	Design effect 4	Design effect 5	Sample size 1	Sample size 2	Sample size 3	Sample size 4	Sample size 5	
<b>OECD</b>	Australia	1.33	2.46	2.94	1.11	3.27	10 919	5 908	4 939	13 062	4 440
	Austria	1.14	3.87	4.26	1.03	4.40	6 171	1 808	1 643	6 791	1 592
	Belgium	1.09	4.60	4.95	1.02	5.04	8 814	2 097	1 951	9 469	1 915
	Canada	1.16	8.79	10.02	1.02	10.17	17 328	2 282	2 003	19 747	1 972
	Chile	1.22	4.43	5.18	1.04	5.40	5 794	1 591	1 362	6 769	1 307
	Czech Republic	1.06	3.69	3.85	1.01	3.91	6 520	1 866	1 791	6 793	1 765
	Denmark	1.11	4.46	4.84	1.02	4.96	6 442	1 606	1 478	7 000	1 445
	Estonia	1.38	2.24	2.72	1.14	3.10	4 043	2 489	2 054	4 899	1 801
	Finland	1.11	3.28	3.53	1.03	3.64	5 308	1 791	1 666	5 707	1 616
	France	1.08	2.30	2.41	1.04	2.49	5 631	2 656	2 534	5 901	2 448
	Germany	1.23	3.92	4.58	1.05	4.80	5 321	1 665	1 425	6 215	1 358
	Greece	1.15	8.70	9.89	1.02	10.05	4 793	636	559	5 447	551
	Hungary	1.12	3.20	3.46	1.03	3.57	5 064	1 769	1 637	5 472	1 583
	Iceland	1.03	1.11	1.11	1.03	1.14	3 266	3 039	3 029	3 276	2 944
	Ireland	1.32	3.14	3.82	1.08	4.14	4 354	1 830	1 504	5 299	1 388
	Israel	1.07	6.42	6.82	1.01	6.89	6 147	1 028	968	6 528	957
	Italy	1.39	6.33	8.38	1.05	8.77	8 359	1 830	1 382	11 074	1 321
	Japan	1.11	6.03	6.58	1.02	6.69	5 989	1 102	1 010	6 538	993
	Korea	1.15	5.23	5.89	1.03	6.04	4 835	1 066	948	5 438	924
	Latvia	1.22	1.44	1.53	1.14	1.75	3 987	3 390	3 177	4 255	2 776
	Luxembourg	1.27	0.52	0.39	1.71	0.66	4 157	10 227	13 738	3 094	8 022
	Mexico	1.44	4.69	6.30	1.07	6.74	5 266	1 614	1 201	7 077	1 123
	Netherlands	1.10	2.46	2.60	1.04	2.69	4 918	2 190	2 073	5 195	2 000
	New Zealand	1.07	2.21	2.30	1.03	2.37	4 206	2 048	1 968	4 378	1 906
	Norway	1.08	2.78	2.93	1.03	3.01	5 037	1 960	1 861	5 306	1 810
	Poland	1.09	3.15	3.33	1.03	3.42	4 125	1 423	1 345	4 366	1 311
	Portugal	1.33	3.87	4.81	1.07	5.13	5 522	1 893	1 524	6 859	1 427
	Puerto Rico (United States) <sup>1</sup>	1.19	5.86	6.78	1.03	6.97	1 175	239	206	1 360	201
	Slovak Republic	1.10	3.98	4.27	1.02	4.36	5 791	1 595	1 488	6 209	1 455
	Slovenia	1.16	1.07	1.08	1.15	1.24	5 503	6 014	5 954	5 558	5 166
	Spain	1.05	3.53	3.66	1.01	3.71	6 418	1 906	1 840	6 646	1 816
	Sweden	1.27	5.30	6.47	1.04	6.74	4 295	1 029	844	5 239	810
	Switzerland	1.15	4.33	4.83	1.03	4.98	5 097	1 354	1 214	5 684	1 178
	Turkey	1.42	10.23	14.10	1.03	14.52	4 152	576	418	5 725	406
	United Kingdom	1.67	5.62	8.71	1.08	9.37	8 484	2 520	1 626	13 147	1 510
	United States	1.18	5.03	5.76	1.03	5.94	4 835	1 135	991	5 538	961
<b>Partners</b>	Albania	1.44	6.33	8.66	1.05	9.10	3 628	824	602	4 964	573
	Algeria	1.48	5.44	7.55	1.06	8.03	3 740	1 014	731	5 192	687
	Argentina	1.65	4.86	7.37	1.09	8.02	3 847	1 306	861	5 834	791
	Brazil	1.40	10.97	14.96	1.03	15.36	16 522	2 110	1 547	22 537	1 507
	B-S-J-G (China)	1.14	17.42	19.66	1.01	19.79	8 661	565	501	9 773	497
	Bulgaria	1.06	10.27	10.82	1.01	10.88	5 596	577	548	5 896	545
	Colombia	1.40	7.29	9.78	1.04	10.18	8 443	1 619	1 206	11 335	1 159
	Costa Rica	1.21	5.00	5.82	1.04	6.03	5 697	1 373	1 179	6 632	1 139
	Croatia	1.12	3.92	4.26	1.03	4.38	5 207	1 480	1 363	5 655	1 327
	Cyprus <sup>2</sup>	1.27	0.97	0.96	1.28	1.23	4 387	5 753	5 804	4 348	4 530
	Dominican Republic	1.59	3.77	5.41	1.11	6.00	2 977	1 258	877	4 272	790
	FYROM	1.30	0.89	0.86	1.35	1.15	4 105	5 981	6 208	3 954	4 611
	Georgia	1.17	3.24	3.62	1.05	3.78	4 553	1 640	1 470	5 081	1 405
	Hong Kong	1.50	3.57	4.85	1.10	5.36	3 569	1 502	1 104	4 857	1 001
	Indonesia	1.56	5.88	8.61	1.06	9.17	4 178	1 107	756	6 116	710
	Jordan	1.28	5.72	7.02	1.04	7.30	5 691	1 271	1 035	6 991	996
	Kazakhstan	1.63	11.10	17.47	1.04	18.10	4 810	706	449	7 568	433
	Kosovo	1.97	1.40	1.78	1.54	2.74	2 455	3 459	2 716	3 126	1 759
	Lebanon	1.32	4.86	6.09	1.05	6.41	3 447	935	746	4 320	709
	Lithuania	1.32	4.20	5.23	1.06	5.55	4 938	1 552	1 247	6 147	1 175
	Macao	1.21	0.63	0.55	1.39	0.77	3 689	7 091	8 101	3 229	5 845
	Malaysia	1.51	9.22	13.43	1.04	13.94	5 860	961	660	8 535	636
	Malta	1.16	0.61	0.55	1.29	0.71	3 140	5 941	6 599	2 827	5 133
	Moldova	1.21	2.31	2.59	1.08	2.80	4 388	2 309	2 060	4 919	1 902
	Montenegro	1.08	0.77	0.75	1.10	0.83	5 255	7 397	7 578	5 129	6 861
	Peru	1.31	5.04	6.28	1.05	6.59	5 326	1 384	1 109	6 644	1 057
	Qatar	1.63	0.77	0.62	2.02	1.25	7 409	15 755	19 491	5 989	9 660
	Romania	1.13	7.22	8.02	1.02	8.15	4 324	675	608	4 800	599
	Russia	1.08	7.02	7.47	1.01	7.55	5 612	860	808	5 976	799
	Singapore	1.12	0.73	0.70	1.17	0.81	5 476	8 379	8 757	5 240	7 504
	Chinese Taipei	1.28	4.40	5.35	1.05	5.63	6 015	1 753	1 440	7 323	1 368
	Thailand	1.36	7.90	10.39	1.03	10.75	6 066	1 044	794	7 973	768
	Trinidad and Tobago	1.39	0.76	0.67	1.59	1.06	3 371	6 167	7 034	2 955	4 430
	Tunisia	1.51	3.75	5.15	1.10	5.66	3 558	1 435	1 044	4 890	950
	United Arab Emirates	1.18	7.17	8.28	1.02	8.46	12 010	1 975	1 711	13 866	1 675
	Uruguay	1.12	3.52	3.81	1.03	3.92	5 436	1 724	1 593	5 884	1 546
	Viet Nam	1.39	10.93	14.79	1.03	15.18	4 195	533	394	5 677	384

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

2. See note 1 under Table 11.1.



Table 11.13 Design effects and effective sample sizes for mathematical literacy

Country	Design effect 1	Design Effect 2	Design effect 3	Design effect 4	Design effect 5	Sample size 1	Sample size 2	Sample size 3	Sample size 4	Sample size 5
<b>OECD</b>	Australia	2.36	1.83	2.96	1.46	4.32	6 157	7 931	4 902	9 960
	Austria	1.77	3.58	5.57	1.14	6.34	3 958	1 958	1 259	6 155
	Belgium	1.38	4.07	5.24	1.07	5.62	6 986	2 370	1 841	8 997
	Canada	2.99	4.66	11.95	1.17	13.94	6 709	4 302	1 678	17 195
	Chile	1.85	3.37	5.38	1.16	6.23	3 818	2 091	1 310	6 094
	Czech Republic	1.63	2.96	4.18	1.15	4.81	4 235	2 332	1 648	5 994
	Denmark	1.76	2.96	4.45	1.17	5.21	4 064	2 420	1 608	6 115
	Estonia	1.86	1.93	2.74	1.32	3.60	2 997	2 890	2 039	4 248
	Finland	2.04	2.28	3.61	1.29	4.65	2 882	2 583	1 631	4 565
	France	1.33	2.23	2.64	1.13	2.97	4 578	2 743	2 316	5 421
	Germany	2.90	2.37	4.97	1.38	6.87	2 249	2 754	1 313	4 717
	Greece	1.96	4.97	8.79	1.11	9.76	2 818	1 113	629	4 986
	Hungary	1.55	2.65	3.56	1.15	4.11	3 651	2 135	1 591	4 901
	Iceland	1.44	1.07	1.10	1.40	1.54	2 336	3 150	3 061	2 404
	Ireland	1.21	3.14	3.59	1.06	3.80	4 736	1 830	1 599	5 421
	Israel	1.93	4.21	7.20	1.13	8.13	3 415	1 567	916	5 842
	Italy	2.53	4.23	9.18	1.17	10.71	4 570	2 741	1 262	9 923
	Japan	2.13	3.61	6.56	1.17	7.69	3 124	1 840	1 014	5 672
	Korea	1.91	4.05	6.82	1.13	7.73	2 919	1 380	818	4 923
	Latvia	1.92	1.48	1.91	1.48	2.83	2 541	3 299	2 547	3 291
	Luxembourg	1.56	0.62	0.41	2.38	0.97	3 389	8 515	12 943	2 229
	Mexico	1.19	5.65	6.56	1.03	6.75	6 336	1 338	1 154	7 350
	Netherlands	1.36	2.31	2.79	1.13	3.14	3 965	2 326	1 933	4 771
	New Zealand	1.36	2.01	2.37	1.15	2.73	3 323	2 249	1 904	3 925
	Norway	1.58	2.39	3.19	1.18	3.77	3 455	2 286	1 710	4 617
	Poland	1.23	2.70	3.10	1.07	3.33	3 635	1 657	1 446	4 166
	Portugal	1.30	3.81	4.65	1.07	4.96	5 623	1 925	1 574	6 878
	Puerto Rico (United States) <sup>1</sup>	1.52	4.80	6.78	1.08	7.30	919	291	206	1 298
	Slovak Republic	1.69	2.92	4.24	1.16	4.93	3 760	2 176	1 498	5 462
	Slovenia	1.23	1.06	1.08	1.22	1.31	5 193	6 031	5 949	5 264
	Spain	1.52	2.86	3.83	1.14	4.34	4 438	2 354	1 761	5 933
	Sweden	1.48	4.58	6.29	1.08	6.77	3 696	1 191	867	5 074
	Switzerland	1.43	3.82	5.03	1.08	5.45	4 109	1 533	1 166	5 402
	Turkey	1.36	10.99	14.61	1.02	14.97	4 328	536	403	5 752
	United Kingdom	1.58	6.50	9.71	1.06	10.30	8 939	2 177	1 457	13 355
	United States	1.45	5.06	6.87	1.07	7.32	3 948	1 129	831	5 363
<b>Partners</b>	Albania	1.38	6.05	7.96	1.05	8.34	3 786	862	655	4 979
	Algeria	1.79	5.32	8.72	1.09	9.50	3 088	1 038	633	5 062
	Argentina	1.33	6.84	8.78	1.04	9.11	4 766	928	723	6 118
	Brazil	5.96	4.01	18.90	1.26	23.86	3 885	5 778	1 224	18 333
	B-S-J-G (China)	2.20	9.48	19.70	1.06	20.90	4 464	1 038	500	9 274
	Bulgaria	1.37	7.16	9.45	1.04	9.82	4 326	828	628	5 704
	Colombia	2.20	4.70	9.15	1.13	10.35	5 359	2 508	1 289	10 426
	Costa Rica	3.30	2.70	6.61	1.35	8.91	2 081	2 543	1 039	5 093
	Croatia	1.84	3.10	4.88	1.17	5.72	3 151	1 872	1 191	4 953
	Cyprus <sup>2</sup>	2.10	0.92	0.83	2.33	1.93	2 654	6 068	6 727	2 394
	Dominican Republic	2.97	2.45	5.31	1.37	7.28	1 595	1 936	893	3 456
	FYROM	1.18	0.81	0.78	1.23	0.95	4 526	6 576	6 861	4 339
	Georgia	1.56	3.00	4.11	1.14	4.67	3 418	1 772	1 293	4 683
	Hong Kong	1.44	4.06	5.42	1.08	5.86	3 719	1 319	990	4 955
	Indonesia	2.08	4.67	8.64	1.13	9.72	3 128	1 395	754	5 788
	Jordan	2.01	3.45	5.93	1.17	6.93	3 617	2 105	1 226	6 210
	Kazakhstan	4.52	4.67	17.60	1.20	21.13	1 733	1 679	445	6 533
	Kosovo	1.41	1.60	1.84	1.22	2.25	3 433	3 020	2 622	3 954
	Lebanon	1.39	4.35	5.67	1.07	6.06	3 266	1 044	802	4 252
	Lithuania	1.38	3.42	4.35	1.09	4.73	4 712	1 910	1 501	5 995
	Macao	1.30	0.66	0.56	1.55	0.86	3 432	6 787	8 051	2 893
	Malaysia	2.26	6.46	13.33	1.09	14.59	3 923	1 372	665	8 096
	Malta	1.27	0.70	0.61	1.43	0.88	2 872	5 227	5 915	2 538
	Moldova	1.65	2.41	3.33	1.20	3.98	3 221	2 211	1 599	4 451
	Montenegro	1.82	0.89	0.79	2.04	1.61	3 109	6 396	7 155	2 779
	Peru	2.52	2.99	6.01	1.25	7.53	2 768	2 332	1 160	5 565
	Qatar	2.44	0.82	0.56	3.56	2.01	4 943	14 710	21 442	3 391
	Romania	1.44	6.52	8.94	1.05	9.38	3 386	748	545	4 647
	Russia	1.62	5.20	7.82	1.08	8.45	3 719	1 160	772	5 591
	Singapore	1.55	0.93	0.89	1.62	1.44	3 936	6 579	6 868	3 771
	Chinese Taipei	2.46	2.71	5.22	1.28	6.68	3 130	2 841	1 477	6 021
	Thailand	1.70	6.74	10.74	1.06	11.43	4 862	1 224	768	7 746
	Trinidad and Tobago	1.45	0.70	0.56	1.79	1.01	3 247	6 728	8 339	2 620
	Tunisia	1.48	4.47	6.13	1.08	6.60	3 639	1 202	878	4 987
	United Arab Emirates	2.43	3.62	7.38	1.19	8.81	5 819	3 914	1 920	11 861
	Uruguay	2.27	2.22	3.77	1.34	5.04	2 671	2 732	1 609	4 534
	Viet Nam	1.54	10.71	15.97	1.03	16.51	3 778	544	365	5 635

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

2. See note 1 under Table 11.1.

Table 11.14 Design effects and effective sample sizes for reading literacy

Country	Design effect 1	Design effect 2	Design effect 3	Design effect 4	Design effect 5	Sample size 1	Sample size 2	Sample size 3	Sample size 4	Sample size 5
<b>OECD</b>										
Australia	2.40	1.64	2.54	1.55	3.94	6 065	8 836	5 712	9 382	3 688
Austria	1.97	2.80	4.54	1.21	5.51	3 558	2 506	1 544	5 774	1 273
Belgium	1.37	4.11	5.24	1.07	5.61	7 069	2 350	1 841	9 022	1 721
Canada	2.56	4.82	10.77	1.14	12.33	7 836	4 163	1 862	17 521	1 626
Chile	1.35	4.48	5.71	1.06	6.06	5 209	1 575	1 235	6 641	1 163
Czech Republic	1.41	3.27	4.20	1.10	4.62	4 882	2 108	1 640	6 279	1 493
Denmark	1.63	3.74	5.46	1.12	6.08	4 398	1 917	1 313	6 421	1 177
Estonia	1.64	2.19	2.95	1.22	3.59	3 413	2 548	1 892	4 596	1 557
Finland	1.12	3.87	4.21	1.03	4.33	5 256	1 521	1 397	5 720	1 359
France	1.34	2.28	2.72	1.13	3.06	4 551	2 680	2 249	5 425	1 997
Germany	1.45	4.08	5.45	1.08	5.90	4 511	1 599	1 197	6 029	1 106
Greece	1.35	8.01	10.45	1.03	10.80	4 103	691	529	5 354	512
Hungary	1.51	2.81	3.74	1.14	4.25	3 744	2 013	1 514	4 977	1 332
Iceland	1.24	1.08	1.10	1.22	1.35	2 708	3 120	3 064	2 757	2 506
Ireland	1.72	2.73	3.98	1.18	4.70	3 341	2 099	1 442	4 863	1 222
Israel	1.18	6.25	7.18	1.02	7.36	5 603	1 056	919	6 439	897
Italy	2.68	3.54	7.79	1.22	9.47	4 326	3 276	1 487	9 530	1 223
Japan	1.41	5.63	7.54	1.05	7.95	4 705	1 181	882	6 302	836
Korea	2.02	3.60	6.26	1.16	7.28	2 760	1 550	892	4 797	767
Latvia	1.39	1.58	1.81	1.21	2.20	3 506	3 074	2 689	4 009	2 214
Luxembourg	1.54	0.63	0.43	2.25	0.97	3 447	8 415	12 302	2 357	5 473
Mexico	2.34	3.55	6.96	1.19	8.30	3 234	2 135	1 088	6 346	912
Netherlands	1.47	2.09	2.60	1.18	3.07	3 659	2 580	2 072	4 558	1 753
New Zealand	1.27	1.87	2.10	1.13	2.37	3 561	2 422	2 153	4 006	1 908
Norway	1.47	2.39	3.05	1.15	3.52	3 707	2 281	1 790	4 725	1 550
Poland	1.58	2.18	2.86	1.20	3.45	2 828	2 058	1 565	3 720	1 300
Portugal	2.01	3.12	5.27	1.19	6.29	3 638	2 346	1 389	6 144	1 165
Puerto Rico (United States) <sup>1</sup>	1.35	5.72	7.36	1.05	7.71	1 038	244	190	1 335	181
Slovak Republic	1.39	3.36	4.29	1.09	4.67	4 569	1 888	1 482	5 821	1 358
Slovenia	1.63	1.01	1.02	1.61	1.65	3 939	6 312	6 255	3 975	3 881
Spain	1.72	2.87	4.21	1.17	4.93	3 917	2 348	1 599	5 753	1 366
Sweden	1.31	4.86	6.08	1.05	6.39	4 153	1 122	898	5 190	854
Switzerland	1.50	3.74	5.12	1.10	5.62	3 896	1 569	1 146	5 334	1 043
Turkey	1.37	9.96	13.27	1.03	13.64	4 302	592	444	5 735	432
United Kingdom	3.06	3.79	9.55	1.22	11.61	4 625	3 732	1 482	11 644	1 219
United States	1.34	4.96	6.31	1.05	6.65	4 262	1 151	905	5 420	859
<b>Partners</b>										
Albania	1.57	6.07	8.96	1.06	9.53	3 322	859	582	4 903	547
Algeria	1.96	4.81	8.45	1.11	9.40	2 823	1 147	653	4 959	587
Argentina	1.27	6.59	8.08	1.03	8.35	5 013	963	786	6 146	761
Brazil	4.68	3.72	13.75	1.27	17.44	4 942	6 214	1 683	18 254	1 327
B-S-J-G (China)	1.41	15.50	21.44	1.02	21.85	6 982	635	459	9 657	450
Bulgaria	1.58	7.12	10.69	1.05	11.27	3 746	832	555	5 622	526
Colombia	2.26	5.60	11.41	1.11	12.67	5 211	2 107	1 034	10 619	931
Costa Rica	1.39	5.48	7.21	1.05	7.59	4 953	1 254	953	6 517	904
Croatia	1.34	3.78	4.72	1.07	5.05	4 341	1 538	1 232	5 420	1 149
Cyprus <sup>2</sup>	1.54	0.95	0.93	1.58	1.47	3 628	5 839	5 994	3 534	3 802
Dominican Republic	1.44	4.26	5.69	1.08	6.13	3 295	1 112	832	4 401	773
FYROM	1.34	0.80	0.74	1.45	1.07	3 988	6 624	7 215	3 662	4 962
Georgia	1.57	2.76	3.76	1.15	4.33	3 395	1 923	1 413	4 621	1 228
Hong Kong	1.37	3.85	4.89	1.07	5.26	3 925	1 391	1 095	4 987	1 019
Indonesia	1.94	4.77	8.32	1.11	9.25	3 359	1 365	783	5 852	704
Jordan	2.01	3.49	6.01	1.17	7.02	3 610	2 083	1 209	6 219	1 035
Kazakhstan	3.52	4.04	11.69	1.22	14.21	2 230	1 941	671	6 452	552
Kosovo	1.34	1.44	1.60	1.21	1.94	3 598	3 342	3 024	3 976	2 491
Lebanon	1.56	4.26	6.08	1.09	6.64	2 918	1 066	747	4 164	684
Lithuania	1.22	4.49	5.26	1.04	5.48	5 350	1 452	1 240	6 264	1 190
Macao	1.54	0.67	0.50	2.08	1.04	2 908	6 638	8 976	2 150	4 312
Malaysia	1.97	8.30	15.36	1.06	16.33	4 503	1 068	577	8 336	543
Malta	1.20	0.66	0.59	1.34	0.79	3 028	5 512	6 149	2 715	4 593
Moldova	1.29	2.74	3.24	1.09	3.53	4 135	1 945	1 645	4 890	1 510
Montenegro	1.81	0.88	0.78	2.04	1.60	3 123	6 441	7 249	2 775	3 550
Peru	1.65	4.43	6.66	1.10	7.31	4 226	1 573	1 046	6 352	953
Qatar	1.44	0.71	0.58	1.76	1.03	8 372	17 008	20 757	6 860	11 785
Romania	1.33	6.71	8.59	1.04	8.93	3 663	727	567	4 695	546
Russia	1.66	4.51	6.84	1.10	7.50	3 630	1 338	883	5 503	805
Singapore	1.72	0.97	0.95	1.76	1.68	3 549	6 282	6 409	3 478	3 646
Chinese Taipei	1.35	4.11	5.18	1.07	5.53	5 723	1 877	1 487	7 225	1 394
Thailand	2.14	6.77	13.37	1.09	14.52	3 849	1 218	617	7 599	568
Trinidad and Tobago	1.29	0.74	0.67	1.43	0.96	3 636	6 316	7 022	3 271	4 895
Tunisia	3.07	2.47	5.52	1.37	7.58	1 752	2 175	974	3 909	709
United Arab Emirates	2.39	4.37	9.05	1.15	10.44	5 925	3 244	1 565	12 280	1 357
Uruguay	1.41	3.00	3.82	1.11	4.22	4 303	2 022	1 589	5 475	1 435
Viet Nam	1.48	10.41	14.91	1.03	15.38	3 942	560	391	5 645	379

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

2. See note 1 under Table 11.1.



Table 11.15 Design effects and effective sample sizes for collaborative problem solving

Country	Design effect 1	Design Effect 2	Design effect 3	Design effect 4	Design effect 5	Sample size 1	Sample size 2	Sample size 3	Sample size 4	Sample size 5
<b>OECD</b>										
Australia	2.74	1.71	2.93	1.59	4.67	5 311	8 513	4 953	9 129	3 112
Austria	1.77	2.67	3.97	1.20	4.74	3 950	2 622	1 767	5 863	1 478
Belgium	1.69	3.36	4.99	1.14	5.68	5 711	2 873	1 935	8 478	1 700
Canada	2.59	3.69	7.97	1.20	9.55	7 749	5 434	2 518	16 723	2 099
Chile	3.04	2.38	5.21	1.39	7.24	2 321	2 958	1 355	5 069	974
Czech Republic	1.75	2.31	3.29	1.23	4.04	3 937	2 986	2 094	5 614	1 705
Denmark	1.81	3.12	4.84	1.17	5.65	3 959	2 294	1 481	6 135	1 268
Estonia	2.39	1.75	2.80	1.50	4.18	2 342	3 186	1 997	3 737	1 335
Finland	1.69	2.20	3.02	1.23	3.71	3 478	2 678	1 945	4 787	1 583
France	2.28	1.56	2.28	1.56	3.56	2 678	3 917	2 684	3 908	1 717
Germany	2.24	2.32	3.95	1.31	5.18	2 918	2 812	1 652	4 968	1 258
Greece	1.56	5.41	7.87	1.07	8.43	3 552	1 022	703	5 166	657
Hungary	1.29	2.67	3.15	1.09	3.43	4 395	2 122	1 799	5 184	1 648
Iceland	1.82	1.06	1.11	1.73	1.93	1 856	3 173	3 028	1 945	1 747
Ireland	1.72	4.65	7.30	1.10	8.02	6 717	2 489	1 587	10 537	1 444
Israel	1.65	4.05	6.02	1.11	6.67	4 032	1 643	1 104	6 001	996
Italy	2.13	2.38	3.94	1.29	5.07	2 617	2 346	1 416	4 335	1 100
Japan	2.25	1.37	1.83	1.68	3.09	2 159	3 556	2 657	2 890	1 577
Korea	1.58	0.75	0.61	1.96	1.19	3 344	7 040	8 715	2 701	4 442
Latvia	2.41	3.03	5.88	1.24	7.29	3 141	2 501	1 287	6 105	1 038
Luxembourg	1.41	2.33	2.88	1.14	3.29	3 815	2 308	1 868	4 712	1 635
Mexico	1.49	1.63	1.93	1.25	2.42	3 036	2 781	2 341	3 607	1 868
Netherlands	1.88	2.09	3.06	1.29	3.94	2 899	2 605	1 783	4 235	1 384
New Zealand	2.15	2.84	4.97	1.23	6.12	3 401	2 577	1 475	5 945	1 197
Norway	1.37	3.04	3.78	1.10	4.15	4 649	2 090	1 678	5 790	1 530
Poland	1.96	1.17	1.33	1.72	2.29	3 275	5 471	4 801	3 731	2 797
Portugal	1.67	2.41	3.35	1.20	4.01	4 036	2 800	2 013	5 614	1 678
Puerto Rico (United States) <sup>1</sup>	1.83	3.65	5.86	1.14	6.69	2 980	1 494	931	4 780	816
Slovak Republic	1.41	8.12	11.07	1.04	11.48	4 166	726	533	5 682	513
Slovenia	2.44	3.92	8.13	1.18	9.57	5 801	3 611	1 742	12 025	1 480
Spain	1.67	3.86	5.78	1.12	6.45	3 412	1 482	988	5 115	885
Sweden	1.31	4.86	6.08	1.05	6.39	4 153	1 122	898	5 190	854
Switzerland	1.50	3.74	5.12	1.10	5.62	3 896	1 569	1 146	5 334	1 043
Turkey	1.37	9.96	13.27	1.03	13.64	4 302	592	444	5 735	432
United Kingdom	3.06	3.79	9.55	1.22	11.61	4 625	3 732	1 482	11 644	1 219
United States	1.34	4.96	6.31	1.05	6.65	4 262	1 151	905	5 420	859
<b>Partners</b>										
Brazil	3.57	4.49	13.45	1.19	16.02	6 491	5 151	1 720	19 435	1 445
B-S-J-G (China)	1.57	10.55	15.96	1.04	16.53	6 280	933	617	9 503	595
Bulgaria	1.27	7.23	8.92	1.03	9.19	4 664	820	665	5 753	645
Colombia	1.25	7.30	8.87	1.03	9.12	9 443	1 616	1 330	11 473	1 294
Costa Rica	2.32	2.88	5.36	1.25	6.68	2 964	2 382	1 281	5 512	1 028
Croatia	1.60	3.03	4.26	1.14	4.87	3 620	1 916	1 363	5 087	1 194
Cyprus <sup>2</sup>	1.91	1.03	1.05	1.86	1.95	2 924	5 429	5 307	2 991	2 850
Hong Kong	1.75	3.25	4.95	1.15	5.70	3 057	1 647	1 082	4 652	940
Lithuania	1.56	3.08	4.24	1.13	4.79	4 190	2 120	1 540	5 766	1 361
Macao	1.35	0.64	0.52	1.67	0.86	3 327	6 978	8 647	2 685	5 187
Malaysia	1.68	9.06	14.52	1.05	15.19	5 281	979	610	8 466	583
Montenegro	1.66	0.88	0.80	1.82	1.45	3 418	6 456	7 108	3 104	3 895
Peru	1.58	3.96	5.69	1.10	6.27	4 408	1 759	1 226	6 324	1 112
Russia	1.67	4.97	7.64	1.09	8.31	3 614	1 213	790	5 549	727
Singapore	1.22	0.79	0.75	1.29	0.96	5 026	7 735	8 206	4 737	6 358
Chinese Taipei	1.84	3.15	4.96	1.17	5.80	4 185	2 449	1 555	6 589	1 329
Thailand	2.16	6.72	13.36	1.09	14.52	3 820	1 227	618	7 590	568
Tunisia	1.57	3.71	5.25	1.11	5.82	3 428	1 449	1 024	4 850	924
United Arab Emirates	2.05	4.52	8.24	1.13	9.29	6 896	3 132	1 720	12 560	1 525
Uruguay	1.38	2.79	3.46	1.11	3.84	4 400	2 173	1 750	5 466	1 578

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

2. See note 1 under Table 11.1.

Table 11.16 Design effects and effective sample sizes for financial literacy

Country	Design effect 1	Design effect 2	Design effect 3	Design effect 4	Design effect 5	Sample size 1	Sample size 2	Sample size 3	Sample size 4	Sample size 5	
<i>OECD</i>	Australia	1.27	2.98	3.52	1.08	3.80	11 426	4 869	4 124	13 490	3 828
	Chile	1.81	3.81	6.10	1.13	6.91	3 887	1 852	1 157	6 222	1 021
	Italy	2.02	4.53	8.15	1.13	9.17	5 722	2 557	1 422	10 289	1 263
	Netherlands	1.58	2.06	2.67	1.22	3.26	3 401	2 618	2 014	4 420	1 653
	Poland	1.48	2.57	3.32	1.14	3.79	3 033	1 743	1 350	3 916	1 181
	Slovak Republic	2.35	2.93	5.52	1.24	6.86	2 708	2 170	1 151	5 105	925
	Spain	1.55	3.63	5.08	1.11	5.64	4 342	1 854	1 325	6 077	1 195
	United States	1.16	5.87	6.64	1.02	6.80	4 930	973	860	5 579	840
<i>Partners</i>	Brazil	3.32	6.52	19.33	1.12	21.65	6 970	3 548	1 197	20 662	1 069
	B-S-J-G (China)	2.07	10.66	21.02	1.05	22.10	4 746	924	468	9 363	445
	Lithuania	1.80	3.42	5.37	1.15	6.18	3 616	1 906	1 214	5 675	1 056
	Peru	1.67	4.12	6.20	1.11	6.87	4 180	1 693	1 124	6 293	1 015
	Russia	1.56	5.25	7.64	1.07	8.21	3 861	1 150	790	5 622	735

To better understand the design effect for a particular country, some information related to the design effects and their respective effective sample sizes are presented in Annex C. In particular, the design effect and the effective sample size depend on:

- **The sample size**, the number of participating schools, the number of participating students and the average within-school sample size, which are provided in Table C.2 (Annex C);
- **The school variance**, school variance estimates and the intraclass correlation, which are provided respectively in Tables C.3 and C.4 (Annex C);
- **The stratification variables**, the intraclass correlation coefficient within explicit strata and the percentage of school variance explained by explicit stratification variables, which are provided respectively in Tables C.5 and C.6 (Annex C).

Finally, the standard errors on the mean performance estimates are provided in Table C.1 (Annex C).

## References

- Cochran, W. (1977), *Sampling Techniques* (3<sup>rd</sup> ed.), John Wiley and Sons.
- OECD (2016a), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264266490-en>.
- OECD (2016b), *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264267510-en>.



---

**12**

## Scaling outcomes

<b>Results of the IRT scaling and population modeling .....</b>	226
<b>Transforming the plausible values to PISA scales.....</b>	233
<b>Linking error .....</b>	237
<b>International characteristics of the item pool .....</b>	237

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.



This chapter illustrates the outcomes of applying the item response theory (IRT) scaling and population model for the generation of plausible values to the PISA 2015 main survey assessment data. In the item response theory (IRT) scaling stage, all available items and data from prior PISA cycles (2006, 2009, 2012) were scaled together with the 2015 data via a concurrent calibration using country-by-language-by-cycle groups. However, only results based on the item parameters for the 2015 items are presented here.

## RESULTS OF THE IRT SCALING AND POPULATION MODELING

The linking design for the PISA main survey was aimed at establishing comparability across countries, languages, assessment modes (paper-based and computer-based assessments), and between the 2015 PISA cycle and previous PISA cycles (as far back as 2006, which had been the last time that science was the major domain). By imposing constraints on the item parameters in the item response scaling, the estimated parameters for trend and new items were placed on the same scale, along with items that were used in previous PISA cycles (but not selected for 2015). An additional outcome of the item response theory scaling is that paper-based (PBA) and computer-based (CBA) assessment items can be placed on the same scale. The items generally fit well across countries, allowing for the use of common international item parameters. These international (or common) parameters are what allow for comparability of results across countries and years. However, there are cases where the international item parameters for a given item do not fit well for a particular country or language group, or subset of countries or language groups. In these instances (i.e. when there is item misfit), which imply interactions in certain groups (e.g. item-by-country/language interactions, item-by-mode interactions, item-by-cycle interactions), item constraints were released to allow the estimation of unique item parameters. This was done for a relatively small number of cases across items and groups.

### Unique item parameter estimation and national item deletion

The item response theory calibration for the PISA 2015 main survey data was carried out separately for each of the PISA 2015 domains (reading, mathematical, science, financial literacy, and collaborative problem solving). Both science (as the main domain in PISA 2015) and collaborative problem solving (CPS) (as a new domain in PISA 2015) included new items; science also included trend items. All of the other domains included trend items only. Item fit was evaluated using the mean deviation and the root mean squared deviation. Both deviations were calculated for all items in each country-language group for each mode and PISA cycle.

The final item parameters were estimated based on a concurrent calibration using the data from PISA 2015 as well as from previous PISA cycles going back to 2006. There were only a few items in mathematics and collaborative problem solving that had to be excluded from the item response theory analyses (in all country-by-language-by-cycle groups) due to either almost no response variance, scoring or technical issues (either problems with the delivery platform or with the coding on the platform), or very low or even negative item total correlations; Table 12.1 gives an overview of these items.

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

Domain	Item	Mode	Reason
Maths (1 item)	CM192Q01	CBA	Technical issue
CPS (4 items)	CC104104 CC104303 CC102208 CC105405	CBA	Very few responses in category 0 Technical issue Very few responses in category 0 Low and negative item-total correlation (correlation close to zero)

Note: The problems observed for the items in the table were shown over all countries.

The international/common item parameters and unique national item parameters were estimated for each domain using unidimensional multigroup item response theory models. For analysis purposes, the international/common item parameters are divided into two groups: scalar invariant and metric invariant parameters. Scalar invariant items correspond to items where the slope and threshold parameters are constrained to be the same in both paper-based and computer-based modes. Metric invariant items correspond to items where the slope is constrained to be the same, but the threshold differs across modes. For new items from science and collaborative problem solving, there are no metric invariant item parameters because these were administered only as part of the computer-based assessment; for financial literacy, all items were constrained to be scalar invariant. As such, only scalar invariant percentages are reported in these domains. For each domain, the scalar and metric invariant item parameters represent the stable linked items between the previous and PISA 2015 scales; the unique parameters are included to reduce measurement error. Table 12.2 shows



the percentage of common and unique item parameters by domain computed by dividing the number of unique item-by-country cells through the total item-by-country cells. Note that the percentage of scalar/metric invariant international/common item parameters was above 90% in cognitive domains with the exception of reading and science. Further, only a small number of items received unique item parameters (either group-specific or the same parameters across a subset of groups) except for reading. In reading, the proportion of scalar/metric invariant international/common item parameters was 89.01%, the proportion of group-specific item parameters was 3.01%, and 7.98% received the same unique item parameters across a subset of countries. For trend items in science, 89.70% received scalar/metric invariant international/common item parameters, while 2.62% received group-specific item parameters, and 7.68% received the same parameters across a subset of countries.

**Table 12.2 Percentage of common and unique item parameters in each domain for PISA 2015**

	Maths	Reading	Science trend	Science new	CPS	Financial literacy
% of unique item parameters (group-specific)	2.16%	3.01%	2.62%	2.05%	1.85%	4.40%
% of unique item parameters (same parameters across a subset of groups)	3.36%	7.98%	7.68%	4.60%	3.19%	2.69%
% of metric invariant common/international item parameters	33.22%	30.33%	20.96%	N/A	N/A	N/A
% of scalar invariant common/international item parameters	61.25%	58.68%	68.74%	93.35%	94.96%	92.91%
Mode and number of items in the PISA 2015 main survey	PBA: 83 items, CBA: 81 items	PBA: 103 items, CBA: 103 items	PBA: 85 items, CBA: 85 items	CBA: 99 items	CBA: 117 items	CBA: 43 items

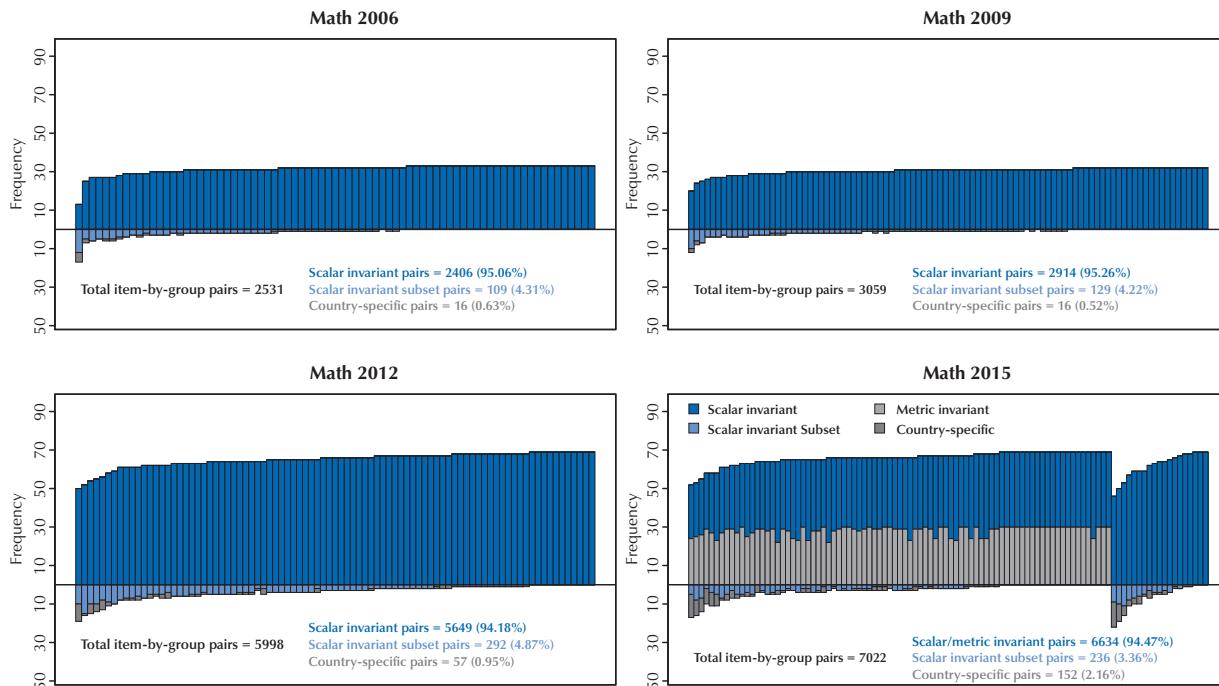
Note: Interactions go across modes and cycles; Kazakhstan is not included due to adjudication issues.

An overview of the proportions of international/common (invariant) item parameters and group-specific item parameters in each domain for each relevant assessment cycle is given in Figures 12.1 to 12.6. The figures also provide an overview of the proportion of scalar invariant item parameters (items sharing common difficulty and slope parameters across modes) and partially or metric invariant item parameters (items sharing common slope parameters across modes) with regard to the mode effect modeling described in Chapter 9: dark blue indicates scalar invariant item parameters, light grey (the lighter grey above the horizontal line) indicates metric invariant item parameters, medium blue indicates scalar invariant item parameters for a subset of groups (unique parameters different from the common parameter, but for several groups sharing the same unique parameter), and dark grey indicates group-specific item parameters. In addition, Annex H provides information about which trend items are scalar invariant and which are partially or metric invariant for each cognitive domain. Recall that both scalar and metric invariant item parameters (dark blue and light grey) contribute to improve the comparability across groups, while unique item parameters (medium blue and dark grey) contribute to the reduction of measurement error. Across every cycle and every domain, it is clear that international/common (invariant) item parameters dominate and only a small proportion of the item parameters are group-specific (i.e. dark grey). Results show that the overall item fit in each domain for each group is very good, resulting in a small numbers of unique item parameters and high comparability of the data. There was no consistent pattern of deviations for any one particular country-by-language group. The results also illustrate that the trend items show good fit, ensuring the quality of the trend measure across different assessment cycles (2015 data versus 2006–2012), different assessment modes (PBA versus CBA), and even across different countries and languages. An overview of the number of deviations per item across all country-by-language-by-cycle groups for items in each domain is given in Annex G.

After the IRT scaling was finalised, item parameter estimates were delivered to each country, including an indication of which items received international/common item parameters and which received unique item parameters. Table 12.3 gives an example of the information provided to countries: the first column shows the domain; the second column shows the flag that indicates whether an item received a unique parameter or was excluded from the IRT scaling; and the remaining columns show the final item parameter estimates (for each item, the slope, difficulty and threshold parameters for polytomous items were listed). A slope parameter of 1 indicates that a Rasch model was fitted for these items; slope estimates different from 1 indicate that the two-parameter logistic model (2PLM) was fitted.

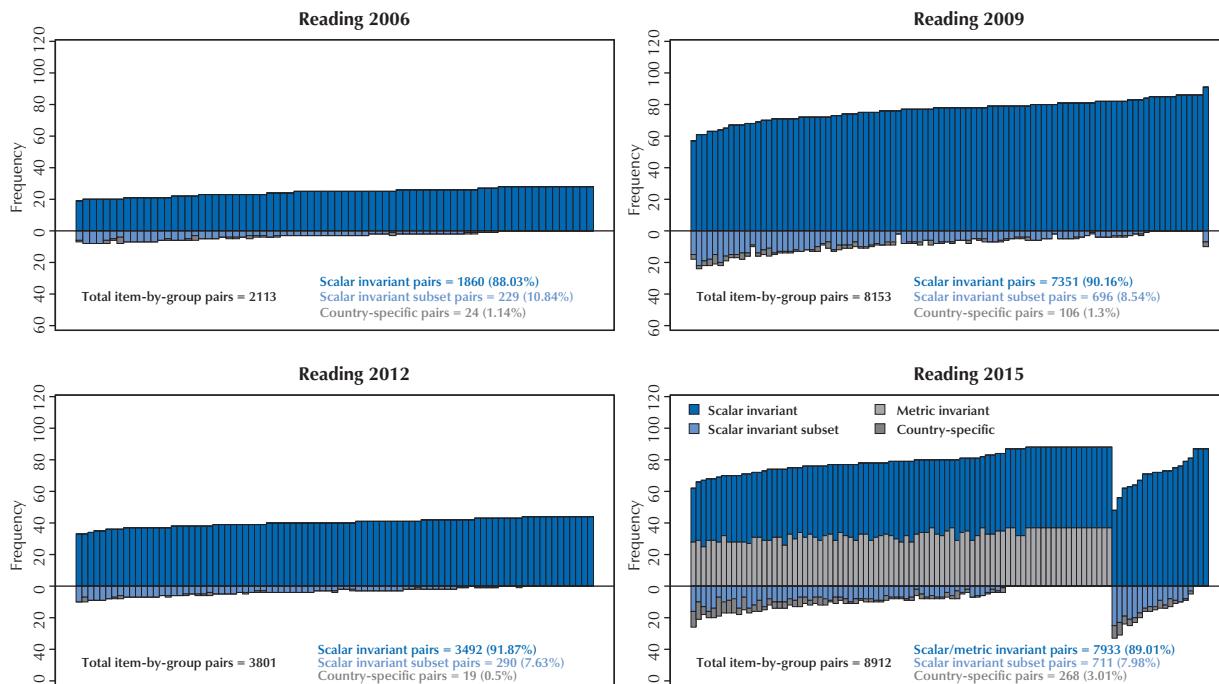
■ Figure 12.1 ■

**Frequencies of international (invariant) and unique item parameters in maths  
(note that frequencies were counted using item-by-group pairs)**



■ Figure 12.2 ■

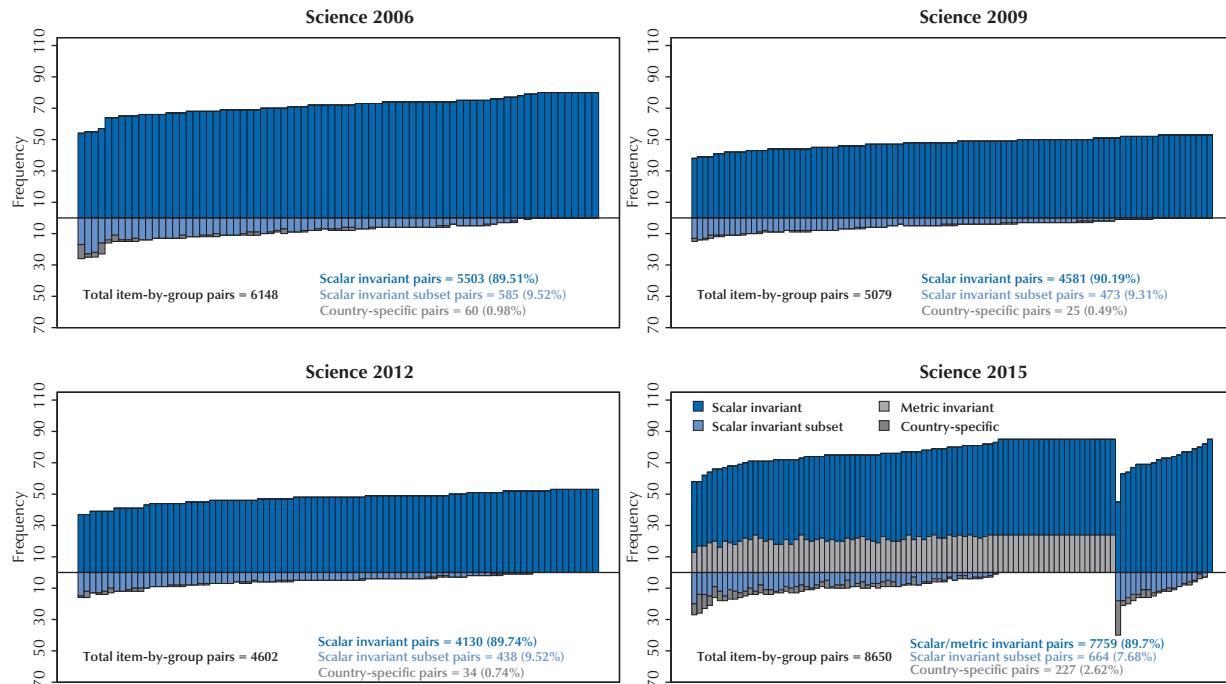
**Frequencies of international (invariant) and unique item parameters in reading  
(note that frequencies were counted using item-by-group pairs)**





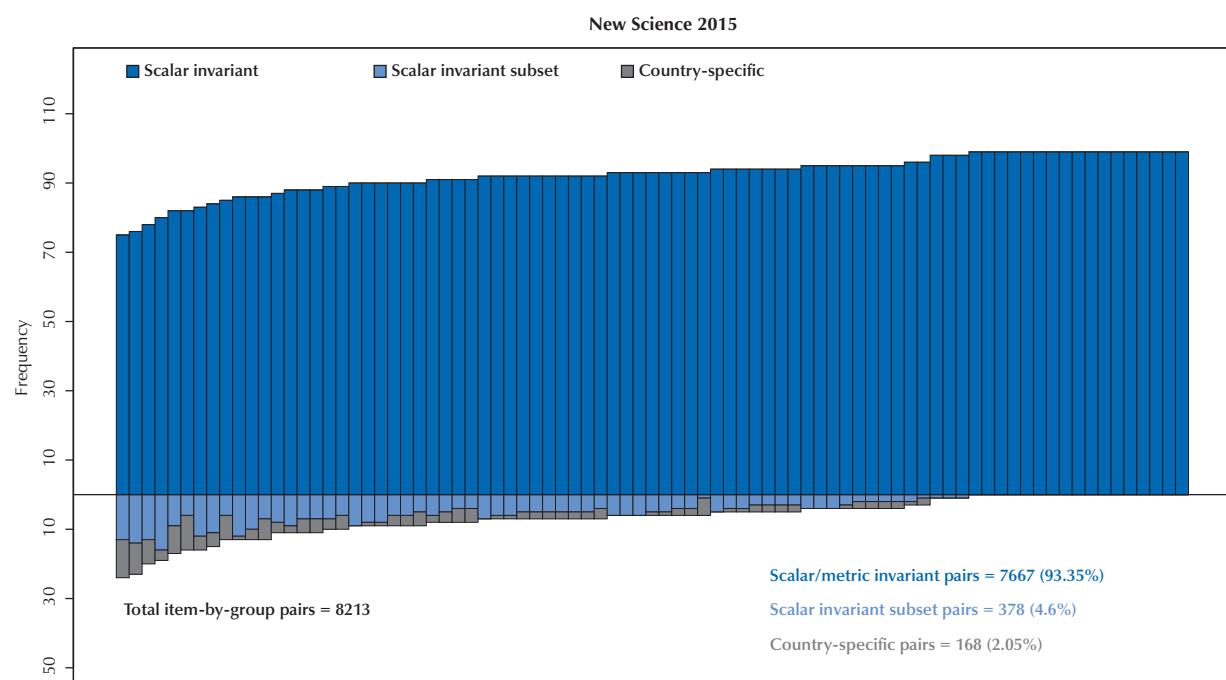
■ Figure 12.3 ■

**Frequencies of international (invariant) and unique item in trend science  
(note that frequencies were counted using item-by-group pairs)**



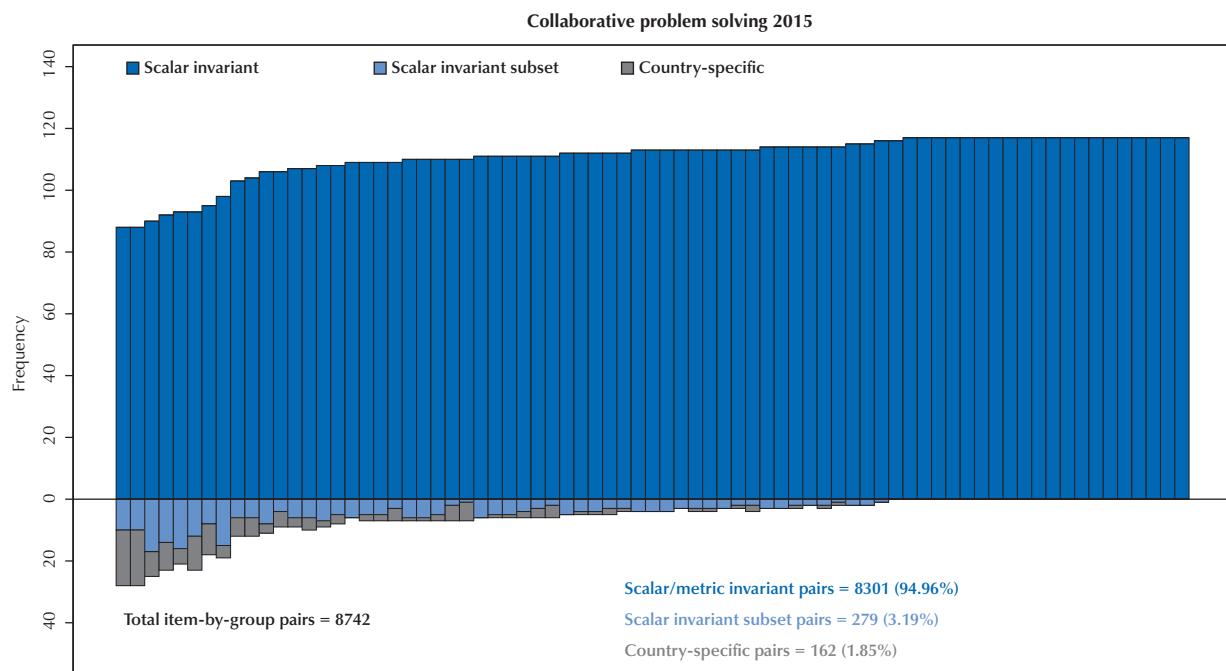
■ Figure 12.4 ■

**Frequencies of international (invariant) and unique item in new science  
(note that frequencies were counted using item-by-group pairs)**



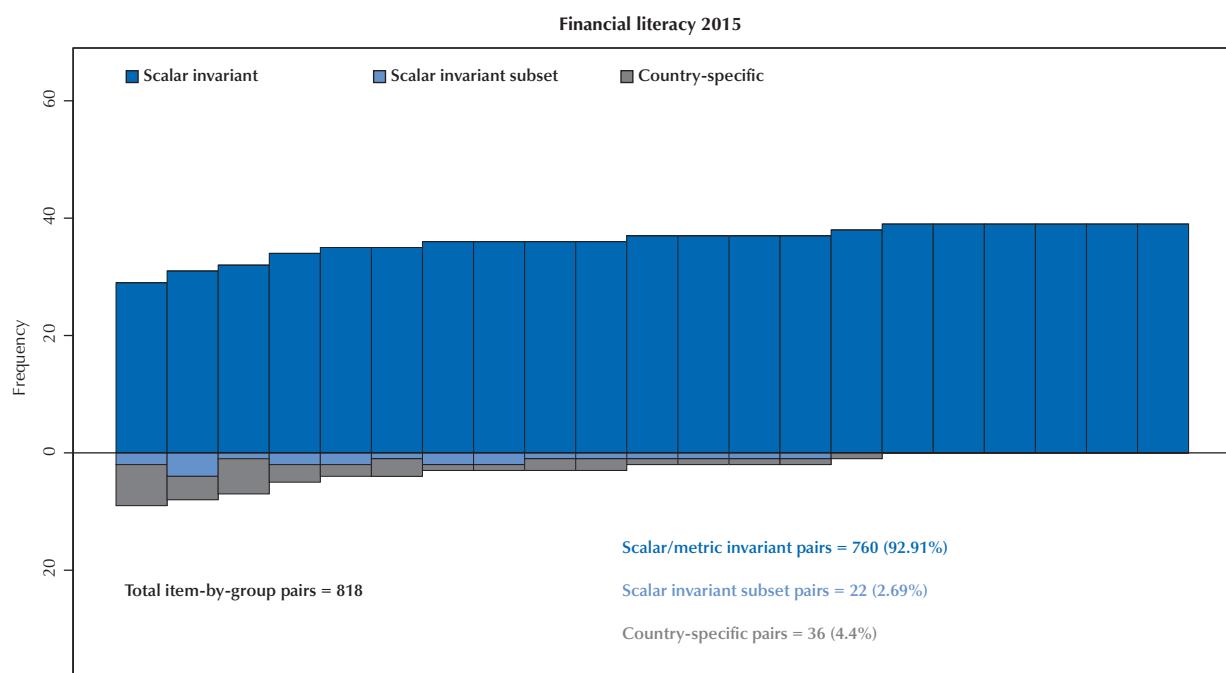
■ Figure 12.5 ■

**Frequencies of international (invariant) and unique item in CPS  
(note that frequencies were counted using item-by-group pairs)**



■ Figure 12.6 ■

**Frequencies of international (invariant) and unique item in financial literacy  
(note that frequencies were counted using item-by-group pairs)**



**Table 12.3 Example of table for item parameter estimates provided to the countries**

Domain	Flag	Item	Slope	Difficulty	IRT_Step1	IRT_Step2
Maths		PM00GQ01	1	1.62226		
Maths		PM00KQ02	1	1.11572		
Maths		PM033Q01S	1	-0.95604		
Maths		PM034Q01S	1	0.15781		
Maths	Unique item parameters	PM155Q01	1.42972	-0.35538		
Maths		PM155Q02	1	-0.35727	-0.42436	0.42436
Maths		PM155Q03	1.08678	0.73497	-0.20119	0.20119
Maths		PM155Q04S	1	-0.27556		
Maths		PM192Q01S	1	0.20948		
Maths	Excluded from scaling	PM936Q01				

### Generating student scale scores and reliability of the PISA scales

Given the rotated and incomplete assessment design, it is not possible to calculate marginal reliabilities for each cognitive domain. In order to get an indication of 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 variance is computed using all 10 plausible values as follows:  $1 - (\text{expected error variance}/\text{total variance})$ . The weighted posterior variance is an expression of the posterior measurement error and is obtained through the 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 posterior variance is the variance across the 10 plausible values). The total variance was estimated using a resampling approach (Efron, 1982). It was estimated for each country depending on the country-specific proficiency distributions for each cognitive domain.

Applying the conditioning approach 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 12.4 gives the median of national reliabilities for the generated scale scores based on all 10 plausible values. National reliabilities of the main cognitive domains based on all 10 plausible values are presented in Table 12.5.

**Table 12.4 Reliabilities of the PISA cognitive domains and Science subscales overall countries<sup>1</sup>**

Mode	Domains	Median	S.D.	Max	Min
CBA	Maths	0.85	0.03	0.90	0.75
	Reading	0.87	0.02	0.90	0.80
	Science	0.91	0.02	0.93	0.82
	CPS	0.78	0.03	0.83	0.70
	Financial literacy	0.83	0.06	0.93	0.72
	<b>Science subscales</b>				
	Explain phenomena scientifically	0.89	0.03	0.91	0.80
	Evaluate and design scientific inquiry	0.87	0.04	0.90	0.71
	Interpret data and evidence scientifically	0.89	0.03	0.92	0.78
	Content	0.89	0.02	0.91	0.81
PBA	Procedural & epistemic	0.90	0.03	0.92	0.78
	Earth & science	0.88	0.03	0.90	0.77
	Living	0.89	0.03	0.91	0.79
	Physical	0.88	0.03	0.91	0.76
	Maths	0.80	0.05	0.87	0.67
	Reading	0.82	0.04	0.88	0.72
	Science	0.86	0.04	0.92	0.77

1. Please note that Argentina, Malaysia, and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

[Part 1/2]

Table 12.5 National reliabilities for main cognitive domains

Mode	Country/economy	Maths	Reading	Science	CPS	Financial literacy
CBA	Australia	0.84	0.86	0.92	0.76	0.93
CBA	Austria	0.87	0.88	0.93	0.80	–
CBA	Belgium	0.89	0.89	0.93	0.80	0.87
CBA	Brazil	0.78	0.82	0.87	0.71	0.72
CBA	Bulgaria	0.85	0.88	0.92	0.82	–
CBA	Canada	0.83	0.83	0.91	0.74	0.76
CBA	Chile	0.86	0.86	0.91	0.78	0.83
CBA	B-S-J-G (China) <sup>1</sup>	0.90	0.90	0.93	0.83	0.88
CBA	Colombia	0.82	0.87	0.89	0.76	–
CBA	Costa Rica	0.78	0.83	0.85	0.70	–
CBA	Croatia	0.86	0.87	0.91	0.76	–
CBA	Cyprus <sup>2</sup>	0.84	0.84	0.90	0.74	–
CBA	Czech Republic	0.88	0.88	0.92	0.77	–
CBA	Denmark	0.85	0.86	0.91	0.78	–
CBA	Dominican Republic	0.81	0.86	0.84	–	–
CBA	Estonia	0.85	0.86	0.91	0.79	–
CBA	Finland	0.85	0.87	0.91	0.77	–
CBA	France	0.88	0.89	0.93	0.77	–
CBA	Germany	0.86	0.86	0.92	0.76	–
CBA	Greece	0.86	0.87	0.91	0.79	–
CBA	Hong Kong (China)	0.84	0.85	0.90	0.77	–
CBA	Hungary	0.88	0.89	0.92	0.81	–
CBA	Iceland	0.83	0.86	0.91	0.76	–
CBA	Ireland	0.85	0.87	0.91	–	–
CBA	Israel	0.87	0.88	0.92	0.83	–
CBA	Italy	0.87	0.87	0.91	0.80	0.81
CBA	Japan	0.85	0.85	0.91	0.75	–
CBA	Korea	0.85	0.85	0.91	0.78	–
CBA	Latvia	0.85	0.86	0.90	0.75	–
CBA	Lithuania	0.85	0.87	0.91	0.80	0.83
CBA	Luxembourg	0.87	0.89	0.93	0.77	–
CBA	Macao (China)	0.82	0.86	0.90	0.78	–
CBA	Malaysia	0.86	0.87	0.90	0.79	–
CBA	Mexico	0.79	0.84	0.86	0.75	–
CBA	Montenegro	0.80	0.84	0.88	0.74	–
CBA	Netherlands	0.89	0.89	0.93	0.79	0.88
CBA	New Zealand	0.85	0.86	0.92	0.79	–
CBA	Norway	0.84	0.85	0.91	0.75	–
CBA	Peru	0.82	0.88	0.87	0.78	0.87
CBA	Poland	0.86	0.87	0.92	–	0.83
CBA	Portugal	0.87	0.86	0.92	0.78	–
CBA	Qatar	0.85	0.89	0.91	–	–
CBA	Russia	0.78	0.80	0.88	0.75	0.73
CBA	Singapore	0.87	0.88	0.93	0.79	–
CBA	Slovak Republic	0.86	0.89	0.92	0.77	0.76
CBA	Slovenia	0.88	0.89	0.93	0.79	–
CBA	Spain	0.86	0.86	0.91	0.75	0.81
CBA	Sweden	0.85	0.86	0.92	0.78	–
CBA	Switzerland	0.86	0.88	0.92	–	–
CBA	Chinese Taipei	0.87	0.88	0.93	0.78	–
CBA	Thailand	0.81	0.86	0.88	0.83	–
CBA	Tunisia	0.75	0.80	0.82	0.70	–
CBA	Turkey	0.82	0.85	0.89	0.74	–
CBA	United Arab Emirates	0.83	0.87	0.91	0.80	–
CBA	United Kingdom	0.87	0.88	0.92	0.83	–
CBA	United States	0.87	0.88	0.92	0.81	0.87
CBA	Uruguay	0.85	0.87	0.90	0.78	–



[Part 2/2]

Table 12.5 National reliabilities for main cognitive domains

Mode	Country/economy	Maths	Reading	Science	CPS	Financial literacy
PBA	Albania	0.75	0.79	0.84	–	–
PBA	Algeria	0.67	0.72	0.77	–	–
PBA	Argentina	0.79	0.82	0.85	–	–
PBA	FYROM	0.79	0.79	0.84	–	–
PBA	Georgia	0.83	0.83	0.86	–	–
PBA	Indonesia	0.78	0.77	0.82	–	–
PBA	Jordan	0.78	0.82	0.86	–	–
PBA	Kazakhstan	0.73	0.71	0.78	–	–
PBA	Kosovo	0.80	0.81	0.82	–	–
PBA	Lebanon	0.82	0.85	0.86	–	–
PBA	Malta	0.87	0.88	0.92	–	–
PBA	Moldova	0.78	0.83	0.86	–	–
PBA	Romania	0.80	0.82	0.86	–	–
PBA	Trinidad and Tobago	0.86	0.84	0.88	–	–
PBA	Viet Nam	0.83	0.84	0.87	–	–

1. B-S-J-G (China) data represent the regions of Beijing, Shanghai, Jiangsu, and Guangdong.

2. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognizes the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue."

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

The table above shows that the explained variance by the combined IRT and latent regression model (population or conditioning model) is at a comparable level across countries. While the population model reaches levels of above 0.80 for reading, mathematics and science, 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 & Sheehan, 1987; von Davier et al., 2006).

## TRANSFORMING THE PLAUSIBLE VALUES TO PISA SCALES

The plausible values were transformed using a linear transformation to form a scale that is linked to the historic PISA scale. This scale can be used to compare the overall performance of countries or subgroups within a country.

For science, reading and mathematics, country results from the 2006, 2009 and 2012 PISA cycles for OECD countries were used to compute the transformation coefficients for each content domain separately. The country means and variances used to compute the transformation coefficients included only those values from the cycle in which a given content domain was the major domain. Hence, the transformation coefficients for science are based on the 2006 reported and model-based results, reading coefficients are based on the 2009 results, and mathematics coefficients are based on the 2012 results. Only the results for countries designated as OECD countries in the respective PISA reporting cycle were used to compute the transformation coefficients. If  $m_{yij}$  is the reported mean for country  $i$  in cycle  $j$ ,  $m_{xij}$  is the model-based mean obtained from the concurrent calibration using the software *mdltm*, and  $s_{yij}^2$  and  $s_{xij}^2$  are the reported and model-based score variances respectively. The same transformation was used for all plausible values (within a given domain). The transformation coefficients for a given content domain were computed as:

**12.1**

$$A = \frac{\tau_{yj}}{\tau_{xj}}$$

**12.2**

$$B = \bar{m}_{yj} - A\bar{m}_{xj}$$



## 12.3

$$\tau_{Yj} = \sqrt{\tau_{Yj}^2} = \sqrt{\left[ \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (m_{Yij} - \bar{m}_{Yj})^2 \right] + \frac{1}{n_j} \sum_{i=1}^{n_j} S_{Yij}^2}$$

## 12.4

$$\tau_{Xj} = \sqrt{\tau_{Xj}^2} = \sqrt{\left[ \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (m_{Xij} - \bar{m}_{Xj})^2 \right] + \frac{1}{n_j} \sum_{i=1}^{n_j} S_{Xij}^2}$$

where =  $\begin{cases} 2006 & \text{Science} \\ 2009 & \text{Reading} \\ 2012 & \text{Maths} \end{cases}$

The values and  $\bar{m}_{Yj}$  and  $\bar{m}_{Xj}$  are grand means of the reported and model-based country means in cycle  $j$ , respectively. The terms  $\tau_{Yj}^2$  and  $\tau_{Xj}^2$  correspond to the total variance, defined as the variance of the country means, plus the mean of the country variances respectively. The square root of these terms is taken to compute the standard deviations  $\tau_{Yj}$  and  $\tau_{Xj}$ . The 2015 plausible values (PVs) for examinee  $k$  in country  $i$  were transformed to the PISA scale via the following transformation:

## 12.5

$$PV_{Tik} = A \times PV_{Uik} + B$$

The subscripts T and U correspond to the transformed and untransformed values respectively.

For financial literacy, country results from the 2012 PISA cycle were used to compute the transformation coefficients. The method used to compute the coefficients is the same as that used for reading, mathematics and science. The key distinction is that in reading, mathematics and science, only results for OECD countries were used to compute the coefficients, whereas, for financial literacy, all available country data were used to compute the coefficients. This decision was made because there were too few OECD countries to provide a defensible transformation of the results. The plausible values for financial literacy were transformed using the same linear transformation as for reading, mathematics and science.

A new scale for CPS was established in PISA 2015. Consistent with the introduction of content domains in previous PISA cycles, transformation coefficients for CPS were computed such that the plausible values for OECD countries have a mean of 500 and a standard deviation of 100. The 10 sets of plausible values were stacked together and the weighted mean and variance (and by extension SD) were computed. Stated differently, the full set of transformed plausible values for CPS have a weighted mean of 500 and a weighted SD of 100 (based on senate weights).

If  $X_{kv}$  is the  $v^{\text{th}}$  PV  $\{v \in 1, 2, \dots, 10\}$  for examinee  $k$ , the transformation coefficients for CPS are computed as

## 12.6

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

## 12.7

$$B = 500 - A[\bar{X}_{kv}] = 500 - A \left[ \frac{\sum_{v=1}^{10} \sum_{k=1}^n X_{kv} W_{kv}}{10 \sum_{k=1}^n W_{kv}} \right]$$

## 12.8

$$\tau_{PV} = \sqrt{\tau_{PV}^2} = \sqrt{\frac{\sum_{v=1}^{10} \sum_{k=1}^n W_{kv} (X_{kv} - \bar{X}_{kv})^2}{[(10n - 1) \sum_{k=1}^n W_{kv}] / n}}$$



The grand mean of the PVs,  $\bar{X}_{kV}$ , was computed by compiling all 10 sets of PVs into a single vector (the corresponding senate weights were compiled in a separate vector) then finding the weighted mean of these values. The weighted variance,  $t_{PV}^2$ , was computed using the vector of PVs as well. The square root is taken to compute the standard deviation,  $\tau_{PV}$ . The plausible values for CPS were transformed using the same approach as that for science, reading, mathematics and financial literacy. The transformations for reading, mathematics, science and financial literacy used the model-based results from the concurrent calibration (IRT scaling) in order to align the results with previously established scales. The transformation for CPS is based on the PVs because this is the first time the results for this domain have been scaled.

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

**Table 12.6 PISA 2015 transformation coefficients**

Domain	A	B
Science	168.3189	494.5360
Reading	131.5806	437.9583
Mathematics	135.9030	514.1848
Financial literacy	140.0807	490.7259
Collaborative problem solving	196.7695	462.8102

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

**[Part 1/2]  
Average plausible values (PVs) and resampling-based standard errors (SE) by country/economy  
for the PISA domains of science, reading, mathematics, financial literacy, and collaborative  
problem solving (CPS)**

**Table 12.7**

Country/economy	Maths		Reading		Science		CPS		Financial literacy	
	Average PV	SE	Average PV	SE						
International average	462	0.32	461	0.34	466	0.31	486	0.36	481	0.95
Albania	413	3.45	405	4.13	427	3.28				
Algeria	360	2.95	350	3.00	376	2.64				
Argentina	409	3.05	425	3.22	432	2.87				
Australia	494	1.61	503	1.69	510	1.54	531	1.91	504	1.91
Austria	497	2.86	485	2.84	495	2.44	509	2.56		
Belgium	507	2.35	499	2.42	502	2.29	501	2.39	541	2.95
Brazil	377	2.86	407	2.75	401	2.30	412	2.30	393	3.84
B-S-J-G (China)	531	4.89	494	5.13	518	4.64	496	3.97	566	6.04
Bulgaria	441	3.95	432	5.00	446	4.35	444	3.85		
Canada	516	2.31	527	2.30	528	2.08	535	2.27	533	4.62
Chile	423	2.54	459	2.58	447	2.38	457	2.69	432	3.74
Colombia	390	2.29	425	2.94	416	2.36	429	2.30		
Costa Rica	400	2.47	427	2.63	420	2.07	441	2.42		
Croatia	464	2.77	487	2.68	475	2.45	473	2.52		
Cyprus <sup>1</sup>	437	1.72	443	1.66	433	1.38	444	1.71		
Czech Republic	492	2.40	487	2.60	493	2.27	499	2.20		
Denmark	511	2.17	500	2.54	502	2.38	520	2.53		
Dominican Republic	328	2.69	358	3.05	332	2.58				
Estonia	520	2.04	519	2.22	534	2.09	535	2.47		
Finland	511	2.31	526	2.55	531	2.39	534	2.55		
France	493	2.10	499	2.51	495	2.06	494	2.42		
FYROM	371	1.28	352	1.41	384	1.25				
Georgia	404	2.78	401	2.96	411	2.42				
Germany	506	2.89	509	3.02	509	2.70	525	2.85		
Greece	454	3.75	467	4.34	455	3.92	459	3.60		
Hong Kong (China)	548	2.98	527	2.69	523	2.55	541	2.95		
Hungary	477	2.53	470	2.66	477	2.42	472	2.35		
Iceland	488	1.99	482	1.98	473	1.68	499	2.26		

Table 12.7

**[Part 2/2]**  
**Average plausible values (PVs) and resampling-based standard errors (SE) by country/economy for the PISA domains of science, reading, mathematics, financial literacy, and collaborative problem solving (CPS)**

Country/economy	Maths		Reading		Science		CPS		Financial literacy	
	Average PV	SE	Average PV	SE						
Indonesia	386	3.08	397	2.87	403	2.57				
Ireland	504	2.05	521	2.47	503	2.39				
Israel	470	3.63	479	3.78	467	3.44	469	3.62		
Italy	490	2.85	485	2.68	481	2.52	478	2.53	483	2.80
Japan	532	3.00	516	3.20	538	2.97	552	2.68		
Jordan	380	2.65	408	2.93	409	2.67				
Kazakhstan	460	4.28	427	3.42	456	3.67				
Korea	524	3.71	517	3.50	516	3.13	538	2.53		
Kosovo	362	1.63	347	1.57	378	1.70				
Latvia	482	1.87	488	1.80	490	1.56	485	2.26		
Lebanon	396	3.69	347	4.41	386	3.40				
Lithuania	478	2.33	472	2.74	475	2.65	467	2.46	449	3.15
Luxembourg	486	1.27	481	1.44	483	1.12	491	1.50		
Macao (China)	544	1.11	509	1.25	529	1.06	534	1.24		
Malaysia	446	3.25	431	3.48	443	3.00	440	3.29		
Malta	479	1.72	447	1.78	465	1.64				
Mexico	408	2.24	423	2.58	416	2.13	433	2.46		
Moldova	420	2.47	416	2.52	428	1.97				
Montenegro	418	1.46	427	1.58	411	1.03	416	1.27		
Netherlands	512	2.21	503	2.41	509	2.26	518	2.39	509	3.32
New Zealand	495	2.27	509	2.40	513	2.38	533	2.45		
Norway	502	2.23	513	2.51	498	2.26	502	2.52		
Peru	387	2.71	398	2.89	397	2.36	418	2.50	403	3.40
Poland	504	2.39	506	2.48	501	2.51			485	2.97
Portugal	492	2.49	498	2.69	501	2.43	498	2.64		
Qatar	402	1.27	402	1.02	418	1.00				
Romania	444	3.79	434	4.07	435	3.23				
Russia	494	3.11	495	3.08	487	2.91	473	3.42	512	3.33
Singapore	564	1.47	535	1.63	556	1.20	561	1.21		
Slovak Republic	475	2.66	453	2.83	461	2.59	463	2.38	445	4.53
Slovenia	510	1.26	505	1.47	513	1.32	502	1.75		
Spain	486	2.15	496	2.36	493	2.07	496	2.15	469	3.19
Sweden	494	3.17	500	3.48	493	3.60	510	3.44		
Switzerland	521	2.92	492	3.03	506	2.90				
Chinese Taipei	542	3.03	497	2.50	532	2.69	527	2.47		
Thailand	415	3.03	409	3.35	421	2.83	436	3.50		
Trinidad and Tobago	417	1.41	427	1.49	425	1.41				
Tunisia	367	2.95	361	3.06	386	2.10	382	1.94		
Turkey	420	4.13	428	3.96	425	3.93	422	3.45		
United Arab Emirates	427	2.41	434	2.87	437	2.42	435	2.43		
United Kingdom	492	2.50	498	2.77	509	2.56	519	2.68		
United States	470	3.17	497	3.41	496	3.18	520	3.64	487	3.80
Uruguay	418	2.50	437	2.55	435	2.20	443	2.29		
Viet Nam	495	4.46	487	3.73	525	3.91				

See note 2 under Table 12.5.



## LINKING ERROR

An evaluation of the magnitude of linking error can be accomplished by considering differences between reported country results from previous PISA cycles and the transformed results from the rescaling. In the application to linking error estimation for the 2015 PISA trend comparisons the robust measure of standard deviation was used, the  $S_n$  statistic (Rousseeuw & Croux, 1993); see Chapter 9 for more information on the linking error approach taken in PISA 2015. The robust estimates of linking error between cycles, by domain are presented in Table 12.8.

The  $S_n$  statistic is available in SAS as well as the R package robustbase. See also <https://cran.r-project.org/web/packages/robustbase/robustbase.pdf>. The  $S_n$  statistic was proposed by Rousseeuw and Croux (1993) as a more efficient alternative to the scaled median absolute deviation from the median (1.4826\*MAD) that is commonly used as a robust estimator of standard deviation.

**Table 12.8 Robust link error (based on absolute pairwise differences statistic  $S_n$ ) for comparisons of performance between PISA 2015 and previous assessments**

Comparison	Maths	Reading	Science	Financial literacy
PISA 2000 to 2015		6.8044		
PISA 2003 to 2015	5.6080	5.3907		
PISA 2006 to 2015	3.5111	6.6064	4.4821	
PISA 2009 to 2015	3.7853	3.4301	4.5016	
PISA 2012 to 2015	3.5462	5.2535	3.9228	5.3309

Note: Comparisons between PISA 2015 scores and previous assessments can only be made to when the subject first became a major domain. As a result, comparisons in mathematics performance between PISA 2015 and PISA 2000 are not possible, nor are comparisons in science performance between PISA 2015 and PISA 2000 or PISA 2003.

## INTERNATIONAL CHARACTERISTICS OF THE ITEM POOL

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

### Test targeting

In addition to identifying the relative discrimination and difficulty of items, IRT can be used to summarise the results for various subpopulations of students. A specific value – the response probability (RP) – can be assigned to each item on a scale according to its discrimination and difficulty, similar to students who receive a specific score along a scale according to their performance on the assessment items (OECD, 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. In PISA, an RP62 value is used for the classification of 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 an item. The RP62 values for all items are presented in Annex A together with the final item parameters obtained from the IRT scaling. The respondent classification into different levels is done by PISA scale scores transformed from the plausible values. Each level is defined by certain score boundaries for each cognitive domain. Tables 12.9 to 12.13 show the score boundaries overall countries used for each cognitive domain along with the percentage of items and respondents classified at each level of proficiency. The decision for the score boundaries for science is explained in Chapter 15; for reading and mathematics the same levels were used that were defined in previous PISA cycles.

**Table 12.9 Item and respondent classification for each score boundary in mathematics**

Level	Score points on the PISA scale	Number of items	Percentage of items	Percentage of respondents
6	Higher than 669.30	27	13.30	1.91
5	Higher than 606.99 and less than or equal to 669.30	23	11.33	6.37
4	Higher than 544.68 and less than or equal to 606.99	50	24.63	13.93
3	Higher than 482.38 and less than or equal to 544.68	41	20.20	20.16
2	Higher than 420.07 and less than or equal to 482.38	39	19.21	21.81
1	Higher than 357.77 and less than or equal to 420.07	12	5.91	18.78
Below 1	Less than 357.77	11	5.42	17.05

**Table 12.10 Item and respondent classification for each score boundary in reading**

Level	Score points on the PISA scale	Number of items	Percentage of items	Percentage of respondents
6	Higher than 698.32	18	7.63	0.70
5	Higher than 625.61 and less than or equal to 698.32	28	11.86	4.96
4	Higher than 552.89 and less than or equal to 625.61	50	21.19	15.45
3	Higher than 480.18 and less than or equal to 552.89	62	26.27	24.14
2	Higher than 407.47 and less than or equal to 480.18	58	24.58	24.36
1a	Higher than 334.75 and less than or equal to 407.47	15	6.36	17.92
1b	262.04 to less than or equal to 334.75	5	2.12	9.12
Below 1b	Less than 262.04	0	0.00	3.34

**Table 12.11 Item and respondent classification for each score boundary in science**

Level	Score points on the PISA scale	Number of items	Percentage of items	Percentage of respondents
6	Higher than 707.93	13	4.45	0.76
5	Higher than 633.33 and less than or equal to 707.93	29	9.93	4.79
4	Higher than 558.73 and less than or equal to 633.33	75	25.68	14.51
3	Higher than 484.14 and less than or equal to 558.73	94	32.19	23.20
2	Higher than 409.54 and less than or equal to 484.14	63	21.58	25.71
1a	Higher than 334.94 and less than or equal to 409.54	15	5.14	20.88
1b	260.54 to less than or equal to 334.94	3	1.03	8.68
Below 1b	Less than 260.54	0	0.00	1.48

**Table 12.12 Item and respondent classification for each score boundary in financial literacy**

Level	Score points on the PISA scale	Number of items	Percentage of items	Percentage of respondents
5	Higher than 624.63	22	26.51	9.36
4	Higher than 549.86 and less than or equal to 624.63	14	16.87	17.38
3	Higher than 475.10 and less than or equal to 549.86	24	28.92	24.31
2	Higher than 400.33 and less than or equal to 475.10	12	14.46	22.63
1	Higher than 325.57 and less than or equal to 400.33	6	7.23	15.73
Below 1	Less than 325.57	5	6.02	10.59

**Table 12.13 Item and respondent classification for each score boundary in CPS**

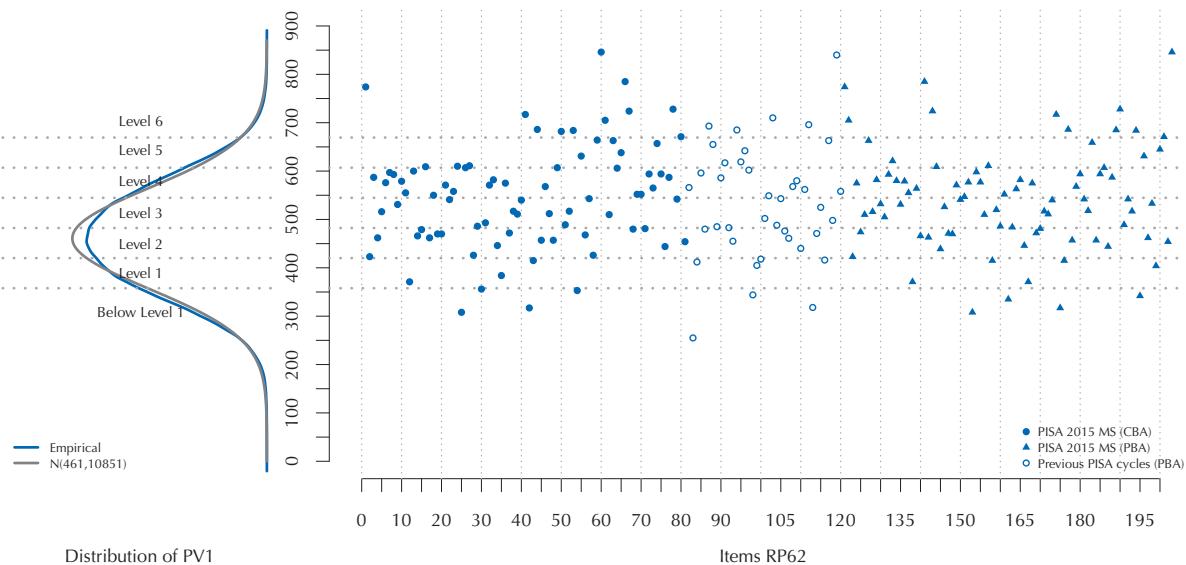
Level	Score points on the PISA scale	Number of items	Percentage of items	Percentage of respondents
4	Higher than 640.00	25	21.37	6.28
3	Higher than 540.00 and less than or equal to 640.00	28	23.93	23.66
2	Higher than 440.00 and less than or equal to 540.00	38	32.48	35.30
1	Higher than 340.00 and less than or equal to 440.00	20	17.09	26.78
Below 1	Less than 340.00	6	5.13	7.99



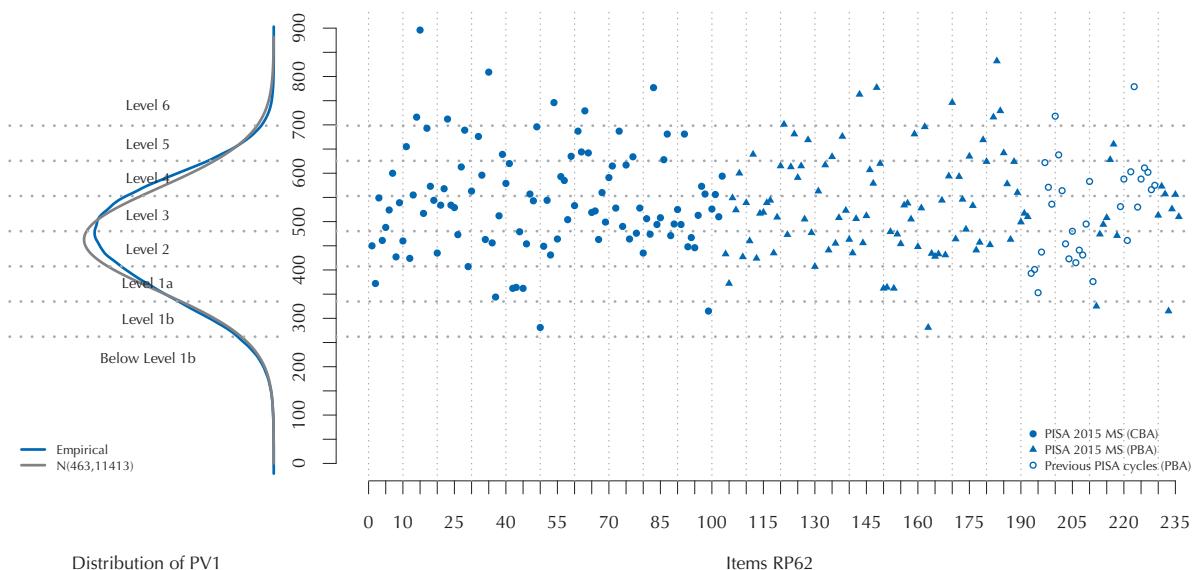
Because RP62 values and the transformed plausible values are on the same PISA scales, the distribution of respondents' latent ability and item RP62 values can be located on the same scale. Figures 12.7 to 12.11 illustrate the distribution of the first plausible value (PV1) along with item RP62 values on the PISA scale separately for each cognitive domain for the PISA 2015 main survey data. Note that international RP62 values and international plausible values (PV1) were used for these figures.<sup>1</sup> RP62 values for CBA items are denoted on the right side. In each domain, solid circles indicate PBA items and hollow circles indicate additional PBA items from previous PISA cycles that were not administered in PISA 2015 main survey. For the polytomous items where partial scoring was available, only the highest RP62 values are illustrated in these figures. 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 grey line indicates the theoretical normal distribution with mean of plausible values and the variance of plausible values in each domain across countries. Specifically,  $N(461, 104.17^2)$  for mathematics,  $N(463, 106.83^2)$  for reading,  $N(467, 103.02^2)$  for science,  $N(474, 123^2)$  for financial literacy, and  $N(483, 101.65^2)$  for CPS are displayed as grey lines. (Note that there are RP62 values higher than 1 000 for the CPS domain, these are outside of the region occupied by the vast majority of respondent's proficiency estimates and therefore are not shown in Figure 12.11.)

■ Figure 12.7 ■

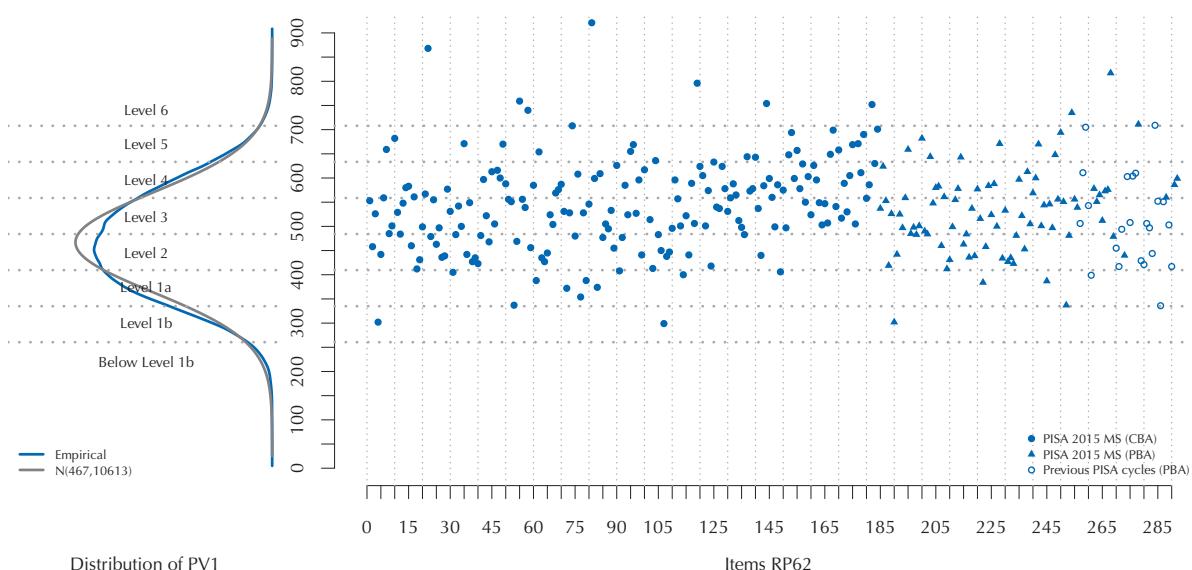
### Item RP62 values and distribution of PV1 in maths



■ Figure 12.8 ■

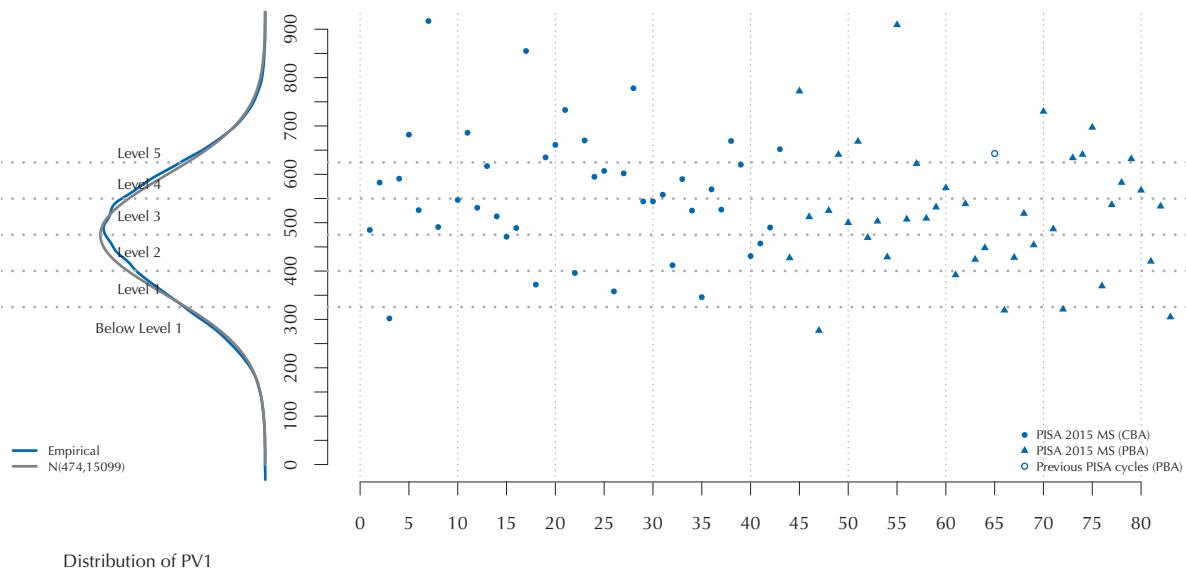
**Item RP62 values and distribution of PV1 in reading**

■ Figure 12.9 ■

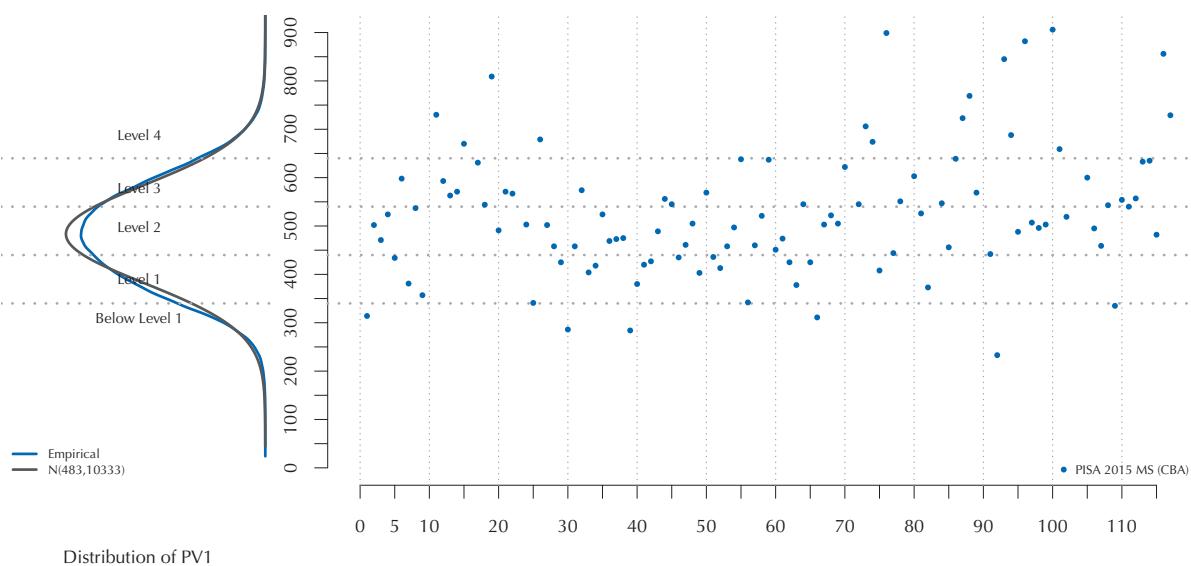
**Item RP62 values and distribution of PV1 in science**



■ Figure 12.10 ■

**Item RP62 values and distribution of PV1 in financial literacy**

■ Figure 12.11 ■

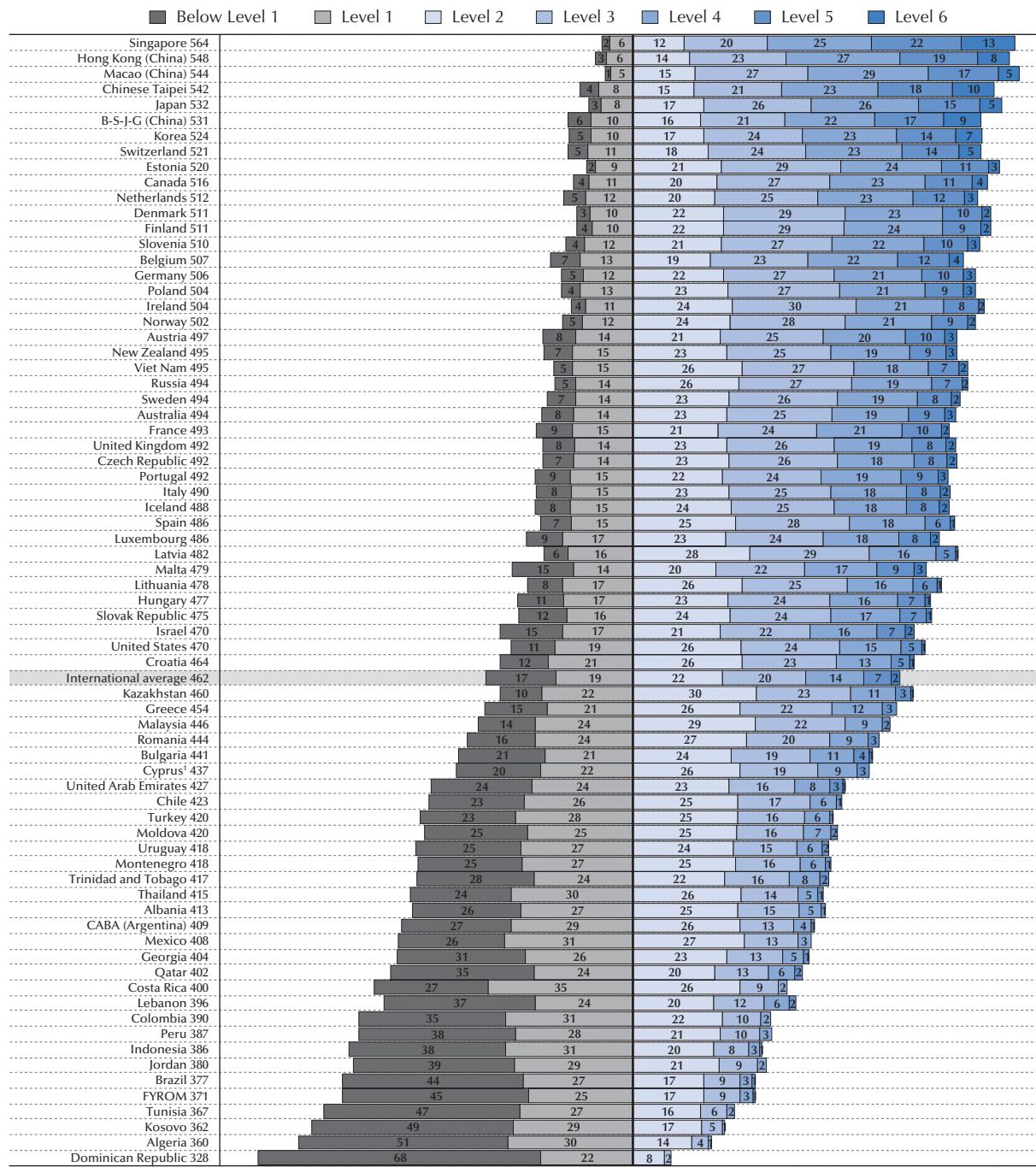
**Item RP62 values and distribution of PV1 in collaborative problem solving**

Figures 12.12 to 12.16 show the percentage of respondents per country at each level of proficiency for each cognitive domain.

■ Figure 12.12 ■

**Percentage of respondents per country/economy at each level of proficiency for maths****2015 PISA main study – maths**

Average scores (PV) &amp; proficiency-level percentages



1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognizes the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue."

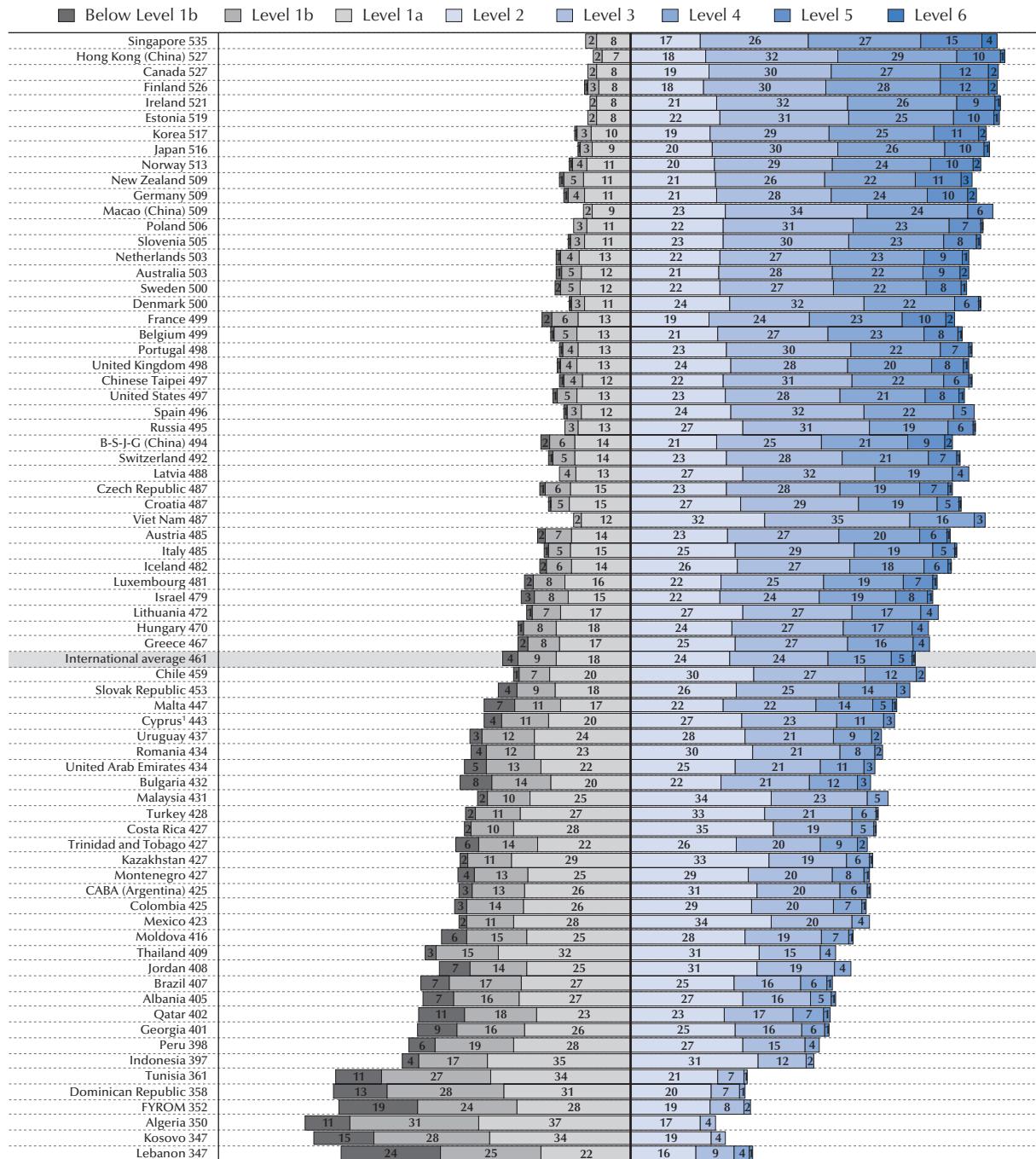
Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.



■ Figure 12.13 ■

### Percentage of respondents per country/economy at each level of proficiency for reading

**2015 PISA main study – reading**  
Average scores (PV) & proficiency-level percentages

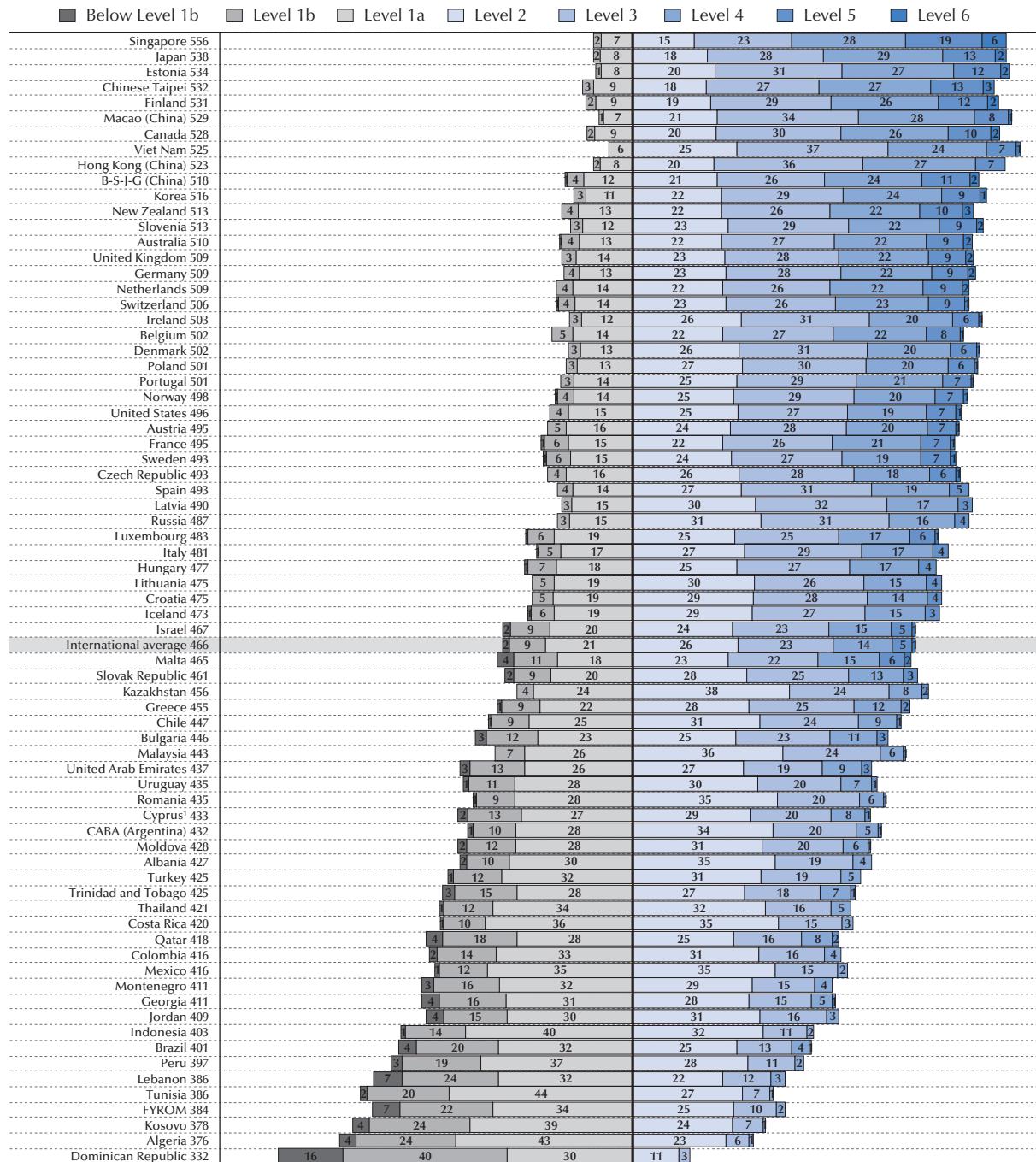


1. See note 2 under Table 12.5.

■ Figure 12.14 ■

**Percentage of respondents per country/economy at each level of proficiency for science**

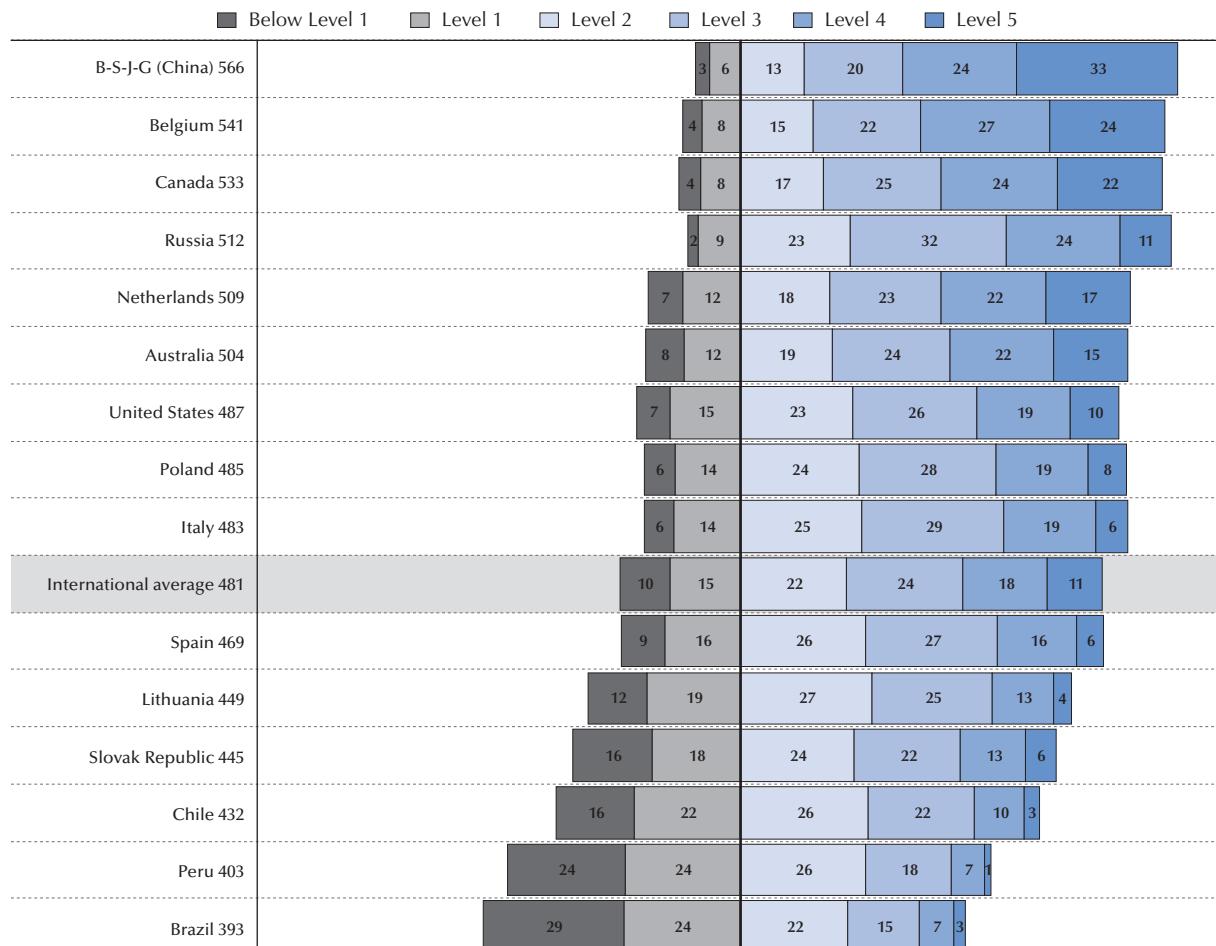
**2015 PISA main study – science**  
Average scores (PV) & proficiency-level percentages



1. See note 2 under Table 12.5.



■ Figure 12.15 ■

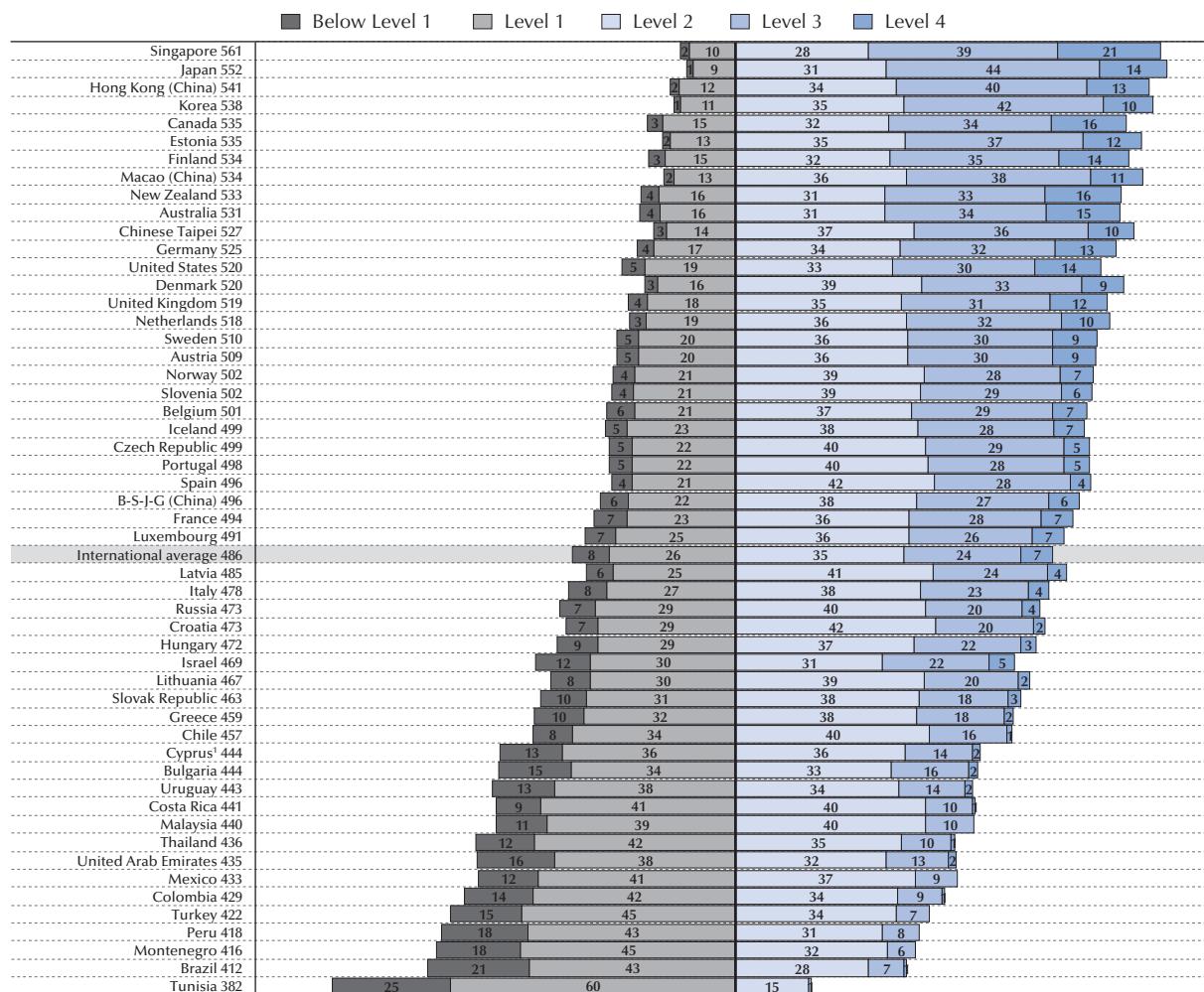
**Percentage of respondents per country/economy at each level of proficiency for financial literacy**
**2015 PISA main study – financial literacy**
*Literacy average scores (PV) & proficiency-level percentages*


Note: The financial literacy data from Belgium come from the Flanders part of Belgium only and thus are not nationally representative; the same is the case with regard to the financial literacy data from Canada since some provinces of Canada did not participate in the financial literacy assessment.

■ Figure 12.16 ■

**Percentage of respondents per country/economy at each level of proficiency for CPS**

**2015 PISA main study – CPS**  
Average scores (PV) & proficiency-level percentages



Note: The CPS sample from Israel does not include ultra-Orthodox students and thus is not nationally representative.

1. See note 2 under Table 12.5.

### Domain inter-correlations

Estimated correlations between the PISA domains, based on the 10 plausible values and averaged across all countries and assessment modes, are presented in Table 12.14. 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 12.15.

**Table 12.14 Domain inter-correlations<sup>1</sup>**

Domain	Reading	Science	CPS	Financial literacy
Maths	Average	0.79	0.88	0.70
	Average (CBA)	0.79	0.88	0.70
	Average (PBA)	0.75	0.80	–
	Range	0.57~0.87	0.70~0.91	0.55~0.76
Reading	Average	–	0.87	0.74
	Average (CBA)		0.87	0.74
	Average (PBA)		0.77	–
	Range		0.71~0.90	0.58~0.80
Science	Average	–	–	0.77
	Average (CBA)			0.77
	Average (PBA)			–
	Range			0.65~0.83
CPS	Average	–	–	0.64
	Average (CBA)			0.64
	Average (PBA)			–
	Range			0.50~0.71

1. Please note that Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

**[Part 1/2]**  
**Table 12.15 National-level domain inter-correlations based on 10 PVs**

Country/economy	Maths & reading	Maths & science	Maths & CPS	Maths & fin. lit.	Reading & science	Reading & CPS	Reading & fin. lit.	Science & CPS	Science & fin. lit.	CPS & fin. lit.
Albania	0.68	0.80	–	–	0.77	–	–	–	–	–
Algeria	0.57	0.70	–	–	0.71	–	–	–	–	–
Argentina	0.75	0.83	–	–	0.81	–	–	–	–	–
Australia	0.79	0.88	0.68	0.79	0.87	0.75	0.8	0.76	0.85	0.7
Austria	0.80	0.89	0.71	–	0.88	0.77	–	0.78	–	–
B-S-J-G (China)	0.84	0.91	0.76	0.80	0.90	0.76	0.80	0.80	0.83	0.70
Belgium	0.84	0.90	0.73	0.80	0.90	0.76	0.80	0.78	0.83	0.67
Brazil	0.75	0.84	0.65	0.62	0.86	0.73	0.65	0.75	0.68	0.54
Bulgaria	0.80	0.89	0.74	–	0.89	0.80	–	0.83	–	–
Canada	0.77	0.87	0.67	0.68	0.87	0.74	0.70	0.75	0.74	0.59
Chile	0.80	0.88	0.70	0.75	0.87	0.74	0.75	0.77	0.78	0.64
Colombia	0.83	0.90	0.74	–	0.90	0.74	–	0.80	–	–
Costa Rica	0.75	0.83	0.59	–	0.85	0.67	–	0.68	–	–
Croatia	0.80	0.89	0.69	–	0.87	0.75	–	0.76	–	–
Cyprus <sup>1</sup>	0.74	0.85	0.65	–	0.83	0.71	–	0.74	–	–
Czech Republic	0.84	0.90	0.69	–	0.89	0.72	–	0.75	–	–
Denmark	0.77	0.87	0.69	–	0.86	0.72	–	0.77	–	–
Dominican Republic	0.78	0.83	–	–	0.85	–	–	–	–	–
Estonia	0.78	0.88	0.71	–	0.87	0.74	–	0.79	–	–
Finland	0.79	0.87	0.72	–	0.87	0.75	–	0.78	–	–
France	0.84	0.91	0.70	–	0.90	0.75	–	0.78	–	–
FYROM	0.75	0.78	–	–	0.74	–	–	–	–	–
Georgia	0.79	0.79	–	–	0.73	–	–	–	–	–
Germany	0.81	0.90	0.70	–	0.88	0.72	–	0.77	–	–
Greece	0.79	0.88	0.73	–	0.88	0.75	–	0.79	–	–
Hong Kong	0.77	0.88	0.64	–	0.86	0.73	–	0.74	–	–
Hungary	0.83	0.90	0.74	–	0.90	0.78	–	0.81	–	–
Iceland	0.78	0.86	0.70	–	0.84	0.74	–	0.76	–	–
Indonesia	0.70	0.82	–	–	0.75	–	–	–	–	–
Ireland	0.81	0.89	–	–	0.88	–	–	–	–	–
Israel	0.83	0.89	0.75	–	0.89	0.78	–	0.80	–	–

[Part 2/2]  
Table 12.15 National-level domain inter-correlations based on 10 PVs

Countries	Maths & Reading	Maths & Science	Maths & CPS	Maths & Fin. Lit.	Reading & Science	Reading & CPS	Reading & Fin. Lit.	Science & CPS	Science & Fin. Lit.	CPS & Fin. Lit.
Italy	0.75	0.85	0.65	0.68	0.84	0.68	0.67	0.73	0.73	0.56
Japan	0.79	0.87	0.66	–	0.86	0.73	–	0.72	–	–
Jordan	0.70	0.79	–	–	0.78	–	–	–	–	–
Kazakhstan	0.61	0.73	–	–	0.70	–	–	–	–	–
Korea	0.78	0.87	0.72	–	0.85	0.76	–	0.77	–	–
Kosovo	0.74	0.81	–	–	0.78	–	–	–	–	–
Latvia	0.77	0.87	0.66	–	0.87	0.73	–	0.75	–	–
Lebanon	0.80	0.82	–	–	0.81	–	–	–	–	–
Lithuania	0.79	0.90	0.72	0.70	0.87	0.74	0.73	0.79	0.75	0.63
Luxembourg	0.83	0.91	0.73	–	0.90	0.78	–	0.78	–	–
Macao	0.75	0.84	0.65	–	0.89	0.78	–	0.78	–	–
Malaysia	0.78	0.87	0.72	–	0.88	0.74	–	0.79	–	–
Malta	0.83	0.87	–	–	0.87	–	–	–	–	–
Mexico	0.77	0.84	0.67	–	0.86	0.73	–	0.76	–	–
Moldova	0.73	0.79	–	–	0.77	–	–	–	–	–
Montenegro	0.76	0.83	0.66	–	0.84	0.70	–	0.74	–	–
Netherlands	0.87	0.91	0.75	0.81	0.89	0.78	0.81	0.77	0.84	0.70
New Zealand	0.79	0.89	0.70	–	0.87	0.75	–	0.78	–	–
Norway	0.78	0.89	0.68	–	0.84	0.72	–	0.74	–	–
Peru	0.81	0.86	0.73	0.76	0.88	0.78	0.81	0.79	0.79	0.70
Poland	0.80	0.90	–	0.74	0.86	–	0.75	–	0.77	–
Portugal	0.79	0.89	0.70	–	0.86	0.74	–	0.76	–	–
Qatar	0.84	0.88	–	–	0.90	–	–	–	–	–
Romania	0.79	0.78	–	–	0.77	–	–	–	–	–
Russian Federation	0.66	0.82	0.55	0.60	0.81	0.68	0.61	0.70	0.68	0.50
Singapore	0.82	0.89	0.73	–	0.90	0.78	–	0.80	–	–
Slovak Republic	0.83	0.88	0.69	0.66	0.87	0.74	0.66	0.74	0.68	0.58
Slovenia	0.79	0.89	0.68	–	0.87	0.73	–	0.74	–	–
Spain	0.76	0.88	0.66	0.71	0.86	0.71	0.72	0.74	0.75	0.61
Sweden	0.78	0.89	0.71	–	0.85	0.78	–	0.77	–	–
Switzerland	0.81	0.88	–	–	0.88	–	–	–	–	–
Chinese Taipei	0.83	0.90	0.71	–	0.90	0.77	–	0.77	–	–
Thailand	0.75	0.83	0.65	–	0.87	0.76	–	0.78	–	–
Trinidad and Tobago	0.81	0.87	–	–	0.80	–	–	–	–	–
Tunisia	0.72	0.81	0.59	–	0.83	0.58	–	0.65	–	–
Turkey	0.76	0.86	0.68	–	0.85	0.71	–	0.76	–	–
United Arab Emirates	0.81	0.88	0.74	–	0.89	0.80	–	0.81	–	–
United Kingdom	0.77	0.87	0.68	–	0.86	0.74	–	0.76	–	–
United States	0.83	0.90	0.76	0.80	0.90	0.79	0.80	0.82	0.83	0.71
Uruguay	0.79	0.88	0.71	–	0.87	0.73	–	0.77	–	–
Viet Nam	0.81	0.87	–	–	0.85	–	–	–	–	–

1. See note 2 under Table 12.5.

### Science scale and subscales

The estimated correlations between the PISA 2015 science subscales and the domains of reading, mathematics, science and financial literacy scales, are presented in Tables 12.16 to 12.18. The different science subscales, which belong to the three scales or subscale groups Knowledge (SKCO, SKPE), Competency (SCEP, SCED, SCID), and System (SSPH, SSLI, SSEs), were considered.

Please note that because of the way in which the proficiency data were generated, you should not calculate the correlations among the knowledge, competency and systems subscales. Therefore these are presented in separate tables.

**Table 12.16 Estimated correlations among domains and science knowledge subscales<sup>1</sup>**

	Reading	Science	CPS	Financial literacy	SKCO	SKPE
Maths	0.783	0.863	0.692	0.726	0.798	0.808
Reading		0.853	0.741	0.738	0.786	0.817
Science			0.765	0.770	–	–
CPS				0.630	0.688	0.722
FinLit					0.743	0.763
SKCO						0.921

Note: Content, SKPE: Procedural & Epistemic.

1. Please note that Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

**Table 12.17 Estimated correlations among domains and science Competency subscales<sup>1</sup>**

	Reading	Science	CPS	Financial literacy	SCED	SCEP	SCID
Maths	0.783	0.863	0.692	0.726	0.778	0.797	0.802
Reading		0.853	0.741	0.738	0.790	0.786	0.805
Science			0.765	0.770	–	–	–
CPS				0.630	0.700	0.687	0.712
FinLit					0.733	0.743	0.756
SCED						0.894	0.903
SCEP							0.919

Note: SCED: Evaluate and Design Scientific Inquiry, SCEP: Subscale of Science Explain Phenomena Scientifically, SCID: Interpret Data and Evidence Scientifically.

1. Please note that Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

**Table 12.18 Estimated correlations among domains and science System subscales<sup>1</sup>**

	Reading	Science	CPS	Financial literacy	SSES	SSLI	SSPH
Maths	0.783	0.863	0.692	0.726	0.791	0.798	0.791
Reading		0.853	0.741	0.738	0.791	0.804	0.781
Science			0.765	0.770	---	---	---
CPS				0.630	0.693	0.711	0.688
FinLit					0.743	0.754	0.736
SSES						0.910	0.900
SSLI							0.908

Note: SSPH: Physical, SSLI: Living, SSES: Earth & Science.

1. Please note that Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).



### Note

1. Please note that Argentina, Malaysia and Kazakhstan were not included in this analysis due to adjudication issues (inadequate coverage of either population or construct).

### References

- Efron, B. (1982), "The Jackknife, the Bootstrap, and Other Resampling Plans", *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, Vol. 38.
- Hoaglin, D.C., F. Mosteller and J.W. Tukey, (1983), *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York, NY.
- 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.
- OECD (2002), *Reading for Change: Performance and Engagement across Countries: Results from PISA 2000*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264099289-en>.
- von Davier et al. (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.



13

# Coding design, coding process, and coder reliability studies

<b>Introduction .....</b>	252
<b>Coding procedures.....</b>	252
<b>Coding preparation .....</b>	253
<b>Coding design .....</b>	254
<b>Coder reliability studies .....</b>	257

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

## INTRODUCTION

The proficiencies of PISA respondents were estimated based on their performance on the test items administered in the assessment. In the PISA 2015 assessment, countries<sup>1</sup> taking part in the computer-based assessment (CBA) administered 18 clusters of trend items from previous cycles – 6 clusters each of mathematics, reading and science – and 6 clusters of new science items developed for 2015. Countries that chose to take part in the financial literacy assessment administered 2 additional clusters of financial literacy items. The tests in countries that used paper-based assessment (PBA) were based solely on the 18 clusters of items from previous PISA cycles.

The PISA 2015 tests consisted of both selected- and constructed-response items. Selected-response items had predefined correct answers that could be computer-coded. While some of the constructed-response items were automatically coded by computer, some elicited a wider variety of responses that could not be categorised in advance, thus requiring human coding. The breakdown of all test items by domain, item format and coding method is shown in Table 13.1.

**Table 13.1 Number of cognitive items by domain, item format and coding method**

Mode	Coding Method	Item Format	Mathematics (trend)	Reading (trend)	Science (trend)	Science (new)	Financial Literacy (trend and new)
CBA	Human	Constructed-response	18 (17)	40 (44)	28	30	16
		Simple selected-response	16 (19)	31 (27)	29	25	10
	Automatic	Complex selected-response	13 (13)	11 (10)	25	41	12
		Constructed-response	22 (20)	6 (6)	3	3	5
<b>Total</b>			<b>69 (69)</b>	<b>88 (87)</b>	<b>85</b>	<b>99</b>	<b>43</b>
PBA	Automatic	Constructed-response	41 (38)	50 (51)	32	NA	
		Simple selected-response	15 (18)	30 (27)	29		
		Complex selected-response	12 (12)	8 (9)	24		
		Constructed-response	3 (3)	0 (0)	0		
<b>Total</b>			<b>71 (71)</b>	<b>88 (87)</b>	<b>85</b>		

Notes: CBA stands for computer-based assessment and PBA stands for paper-based assessment

Consistent with previous cycles, easier and standard forms were developed for mathematics and literacy. Number in the cell corresponds to the standard forms while the number in parenthesis corresponds to the easier form.

New science and financial literacy are CBA domains only.

The six parts of the trend Reading unit, Employment, R219, were separately coded to achieve consistent and accurate scoring. Note that, in the final item counts, four parts related to completing an employment application form were counted as a single item.

The multiple coding design in PISA 2015 included all human-coded items for monitoring coder reliabilities within countries as well as across countries. This chapter aims to describe coding procedures and preparation, coding design options and coding reliability studies.

## CODING PROCEDURES

For computer-based assessment (CBA) participants, the coding designs for the CBA responses for mathematics, reading, science and financial literacy (when applicable) were greatly simplified through use of the open-ended coding system (OECS). This computer system, developed for PISA 2015, supported coders in their work to code the CBA responses while ensuring that the coding design was appropriately implemented. Detailed information about the system was included in the OECS manual. The OECS system worked offline, meaning coders did not need a network connection. It organised responses according to the agreed-upon coding designs.

During the CBA coding, coders worked only with individual PDF files, one for each item, containing one page per item response to be coded. Each page displayed the item stem or question, the individual response and the available codes for the item. The coder was instructed to click the circle next to the selected code, which was then saved within the file. Also included on each page were two checkboxes labelled “recoded” and “defer”. The recoded box was checked if the response had been recoded by another coder for any reason. The defer box was used if the coder was not sure what code to assign to the response. These deferred responses were later reviewed and coded by the coder. It was expected that coders would code the majority of responses for which they were responsible and defer responses only in unusual circumstances. When deferring a response, it was suggested that the coder enter comments into the box labeled “comment” to indicate the reason for deferring the given response. Coders worked on one file until all responses in that file were coded. The process was repeated until all items were coded. The approach of coding by item has been shown to improve reliability and was greatly facilitated by the open-ended coding system (OECS).

For paper-based assessment (PBA) participants, the coding designs for the PBA responses for mathematics, reading and science were supported by the data management expert system and reliability was monitored through the open-ended



reporting system (OERS), a computer tool that worked in conjunction with the data management expert (DME) software to evaluate and report reliability for paper-based, open-constructed responses. Detailed information about the system was provided in the OERS Manual. The coding process for PBA participants involved using the actual paper booklets, with some booklets single coded and others multiple coded by two or more coders. When single coded, coders marked directly in the booklets. When multiple coded, coders coded first on the coding sheets, while the last coder coded directly in the booklet.

National Centres used the output reports generated by the OECS and OERS to monitor irregularities and deviations in the coding process. Careful monitoring of coding reliability plays an important role in data quality control. Through coder reliability monitoring, coding inconsistencies or problems within and across countries could be detected early in the coding process through OECS/OERS output reports, allowing action to be taken as soon as possible. The OECS/OERS worked in concert with the DME database to generate two types of reliability reports: i) proportion agreement and ii) coding category distributions. National Project Managers (NPMs) were instructed to investigate whether a systematic pattern of irregularities existed and was attributable to a particular coder or item. In addition, they were instructed not to carry out resolution (e.g. changing coding on individual responses to reach higher coding consistency). Instead, if systematic irregularities were identified, all responses from a particular item or a particular coder needed to be recoded, including those that showed disagreement as well as those that showed agreement. In general, inconsistencies or problems were due to misunderstanding of general scoring guidelines and/or a rubric for a particular item or misuse of OECS/OERS. Coder reliability studies also made use of the OECS/OERS reports submitted by National Centres.

## CODING PREPARATION

Prior to the assessment, a number of key activities were completed by National Centres to prepare for the process of coding responses to the human-coded constructed-response items.

### Recruitment of national coder teams

National Project Managers were responsible for assembling a team of coders. Their first task was to identify a lead coder who would be part of the coding team and additionally be responsible for the following tasks:

- training coders within the country
- organising all materials and distributing them to coders
- monitoring the coding process
- monitoring the inter-rater reliability and taking action when the coding results were unacceptable and required further investigation
- retraining or replacing coders if necessary
- consulting with the international experts if item-specific issues arose
- producing reliability reports.

The lead coder was required to be proficient in English (as international training and interactions with the contractors were in English only) and to attend the international coder trainings in Malta in January 2014 and Portugal in January 2015. It was also assumed that the lead coder for the field trial would retain the role for the main survey. When this was not the case, it was the responsibility of the National Centre to ensure that the new lead coder received training equivalent to that provided at the international coder training prior to the field trial.

The guidelines for assembling the rest of the coding team included the following requirements:

- All coders should have more than a secondary qualification (i.e., high school degree); university graduates were preferable.
- All should have a good understanding of secondary level studies in the relevant domains.
- All should be available for the duration of the coding period, which was expected to last two to three weeks.
- Due to normal attrition rates and unforeseen absences, it was strongly recommended that lead coders train a backup coder for their teams.
- Two coders for each domain must be bilingual in English and the language of the assessment.



## International coder training

Detailed coding guides were developed for all the new science items that included coding rubrics as well as examples of correct and incorrect responses. For trend items, coding information from previous cycles was included in the coding guides. For new items, coding rubrics were defined for the field trial and then information from field trial coding was used to revise the coding guides for the main survey.

Prior to the field trial, National Project Managers (NPMs) and lead coders were provided with a full item-by-item coder training in Malta in January 2014. The field trial training covered all the items across all domains. Prior to the main survey, NPMs and lead coders were provided with a new round of full item-by-item coder training in Portugal in January 2015. The main survey training covered all new items as well as a set of trend science and trend reading items that required additional training based on the field trial experience. During these trainings, the coding guides were presented and explained. Training participants practiced coding on sample items and discussed any ambiguous or problematic situations as a group. By focusing on sample responses most challenging to code, training participants had the opportunity to ask questions and get the coding rubrics clarified as much as possible. When the discussion revealed areas where rubrics could be improved, those changes were made and included in an updated version of the coding guide documents available after the meeting. As in previous cycles, a “workshop” version of the coding guides was also prepared for the national training. This version included a more extensive set of sample responses; the official coding for each response and a rationale for why each response was coded as shown.

To support the national teams during their coding process, a coder query service was offered. This allowed national teams to submit coding questions and receive responses from the relevant domain experts. National teams were also able to review questions submitted by other countries along with the responses from the test developers. In the case of trend items, responses to queries from previous cycles were also provided. A summary report of coding issues was provided on a regular basis and all related materials were archived in the PISA 2015 portal for reference by national coding teams.

## National coder training provided by the National Centres

Each National Centre was required to develop a training package for their own coders. The training package consisted of an overview of the survey and their own training manuals based on the manuals and materials provided by the international PISA contractors. Coding teams were asked to work on the same schedule and at the same location in order to facilitate discussion about any items that proved challenging. Past experience has shown that if coders can discuss items among themselves and with their lead coder, many issues can be resolved in a way that results in more consistent coding. Each coder was assigned a unique coder ID that was specific to each domain and design.

The National Centres were responsible for organising training and coding using one of the following two approaches and checking with contractors in the case of deviations:

- Coder training took place at the “item” level. Under this approach, coders were fully trained on coding rules for each item and proceeded with coding all responses for that item. Once that item was done, training was provided for the next item and so on.
- Coder training took place at the “item set” level. While coding was conducted at the “item” level, the coder training took place at the “item set” level, with each “item set” containing a few units of items. In this alternative approach, coders were fully trained on a set that varied from 13 to 18 items. Once the full training was complete, coding took place at the item level. However, to ensure that the coding rules were still fresh in the coders’ memory, a coding refresher was recommended before the coding of each item.

## CODING DESIGN

In order to meet the unique characteristics of the CBA participants during the main survey while ensuring that the coding process was completed within a two-to-three week period, ten possible coding designs (one standard design and nine variations) were offered to the CBA participants and four possible coding designs (one standard design and three variations) were offered to the PBA participants.<sup>2</sup> Those designs were developed to accommodate participants’ various needs in terms of the number of languages assessed, the sample size and the specified number of coders required in each domain.

The number of coders by domain in each CBA coding design is shown in Table 13.2. The design of multiple coding in the CBA standard coding design is shown in Table 13.3. In CBA coding designs, human-coded items were bundled into one item set or multiple item sets in each domain. For each common item, coders coded a set of 100 student responses that were randomly selected from all the student responses. Each domain had two bilingual coders who needed to code an additional ten anchor responses for each item assigned to both of them. The rest of the student responses to each item were evenly split



among coders to be single coded. The difference in multiple coding between the standard coding design and other CBA coding designs mainly lay in the number of coders in each domain and which item sets were assigned to each coder.

**Table 13.2 Number of CBA coders by domain and coding design**

Design label	Sample size requirements	Mathematics (trend)	Reading (trend)	Science (trend and new)	Financial literacy (trend and new)
Standard design	Countries with the standard sample size (4 000 – 7 000) for a given language	4	6	8	4
Alternative design 1	Countries with a sample between 7 000 and 9 000 for a given language	4	9	12	4
Alternative design 1a	Countries with a sample between 7 000 and 9 000 for a given language	16	9	12	16
Alternative design 2	Countries with a sample between 9 000 and 13 000 for a given language	6	9	16	6
Alternative design 2a	Countries with a sample between 9 000 and 13 000 for a given language	6	12	16	6
Alternative design 3	Countries with a sample between 13 000 and 19 000 for a given language	6	12	20	6
Alternative design 3a	Countries with a sample between 13 000 and 19 000 for a given language	12	27	32	12
Alternative design 4	Countries with a sample larger than 19 000 for the majority language	9	21	36	6
Minority Language Design 1	Countries with a sample less than 1 500 for the minority language	2	2	2	2
Minority Language Design 2	Countries with a sample between 1 500 and 4 000 for the minority language	3	3	4	3

**Table 13.3 Multiple coding in CBA standard coding design**

		Coder IDs					
Mathematics (trend)	Number of responses for multiple coding	301 (bilingual)	302	303 (bilingual)	304	305	306
Item Set 1	100 student responses per item	✓	✓	✓	✓		
Item Set 1	10 anchor responses per item	◆		◆			
Reading (trend)	Number of responses for multiple coding	201 (bilingual)	202	203 (bilingual)	204	205	206
Item Set 1	100 student responses per item	✓	✓	✓		✓	✓
Item Set 2	100 student responses per item	✓	✓	✓	✓		
Item Set 3	100 student responses per item			✓	✓	✓	✓
Item Set 1	10 anchor responses per item	◆					
Item Set 2	10 anchor responses per item	◆		◆			
Item Set 3	10 anchor responses per item			◆			
Science (trend and new)	Number of responses for multiple coding	101 (bilingual)	102	103 (bilingual)	104	105	106
Item Set 1	100 student responses per item	✓	✓				✓
Item Set 2	100 student responses per item	✓	✓		✓	✓	
Item Set 3	100 student responses per item			✓	✓	✓	✓
Item Set 4	100 student responses per item			✓	✓		✓
Item Set 1	10 anchor responses per item	◆					
Item Set 2	10 anchor responses per item	◆					
Item Set 3	10 anchor responses per item			◆			
Item Set 4	10 anchor responses per item			◆			
Financial Literacy (trend and new)	Number of responses for multiple coding	401 (bilingual)	402	403 (bilingual)	404	405	406
Item Set 1	100 student responses per item	✓	✓	✓	✓		
Item Set 1	10 anchor responses per item	◆		◆			

Note: “✓” denotes the coder should code 100 student responses for each item in the item set. “◆” denotes the coder should code 10 anchor responses for each item in the item set.

Four variations of coding design were offered to PBA participants (See Table 13.4). The design of multiple coding in the PBA standard coding design is shown in Table 13.5. For PBA participants, all paper-and-pencil booklets were organised by form type into 27 different bundle sets: 9 bundle sets per domain. Bundle sets 1, 2 and 3 in each domain were composed of forms for multiple coding: forms 13, 15 and 17 for mathematics; forms 1, 3 and 5 for reading; and forms 7, 8 and 9 for science. For each form, 100 student booklets were randomly selected from all the student responses. Each coder coded his or her assigned clusters on the sets of 100 student booklets until all items in the booklets were coded. Bundle sets 4-9 in each domain were composed of 6 or 7 types<sup>3</sup> of anchor forms. The forms were labelled 301-307 for mathematics; 201-207 for reading; and 101-106 for science (see Table 13.5). Differing from non-anchor forms, the anchor forms each contained only one cluster of items. For example, form 301 contained all the items from the first cluster in maths and form 202 contained all the items from the second cluster in reading. Each anchor form had 10 pre-filled English booklets that were coded by the bilingual coders from each domain. Each domain in the PBA standard design had two bilingual coders: 31 and 33 for mathematics, 21 and 23 for reading and 11 and 13 for science.

CBA constructed-response items were organised by item set during multiple coding; by contrast, PBA constructed-response items were organised by bundle set during multiple coding. In other words, multiple coding in the PBA standard design was form- rather than item-set-based. Although coders conducted coding on the booklets, each coder only coded the clusters assigned to him or her for each booklet, leaving the rest of the clusters to other coders. This multiple coding design enabled the within- and across-country comparison. After the multiple coding was completed, all the clusters that remained uncoded were equally split among coders and coded only once. The difference in multiple coding between the PBA standard design and other PBA coding designs mainly lay in the number of coders in each domain and which forms were assigned to each coder.

**Table 13.4 Number of PBA coders by domain and coding design**

Design Label	Sample Size Requirements	Mathematics (trend)	Reading (trend)	Science (trend and new)
Standard design	Countries with the standard sample size (3,501 – 5,500)	4	6	6
Alternative design 1	Countries with a sample larger than 5,500 for the majority language	6	9	9
Minority language design 1	Countries a sample less than 1,500 for the minority language	2	2	2
Minority design 2	Countries with a sample between 1,501 and 3,500 for the minority language	3	3	4

**Table 13.5 Multiple coding in PBA standard coding design**

Mathematics (trend)	Forms (Clusters)	Number of Booklets per Form	Coder IDs			
			31 (bilingual)	32	33 (bilingual)	34
Bundle set 1	Form 13 (PM1&PM2)	100 student booklets	✓	✓	✓	✓
Bundle set 2	Form 15 (PM3&PM4)	100 student booklets	✓	✓	✓	✓
Bundle set 3	Form 17 (PM5&PM6a or PM5&PM6b )	100 student booklets	✓	✓	✓	✓
Bundle set 4	Form 301 (PM1)	10 anchor booklets	◆	◆		
Bundle set 5	Form 302 (PM2)	10 anchor booklets	◆	◆		
Bundle set 6	Form 303 (PM3)	10 anchor booklets	◆	◆		
Bundle set 7	Form 304 (PM4)	10 anchor booklets	◆	◆		
Bundle set 8	Form 305 (PM5)	10 anchor booklets	◆	◆		
Bundle set 9	Form 306 (PM6a) or 307 (PM6b)	10 anchor booklets	◆	◆		
Reading (trend)	Forms (Clusters)	Number of Booklets per Form	Coder IDs			
			21 (bilingual)	22	23 (bilingual)	24
Bundle set 1	Form 1 (PR1&PR2)	100 student booklets	✓	✓		✓
Bundle set 2	Form 3 (PR3&PR4)	100 student booklets	✓	✓	✓	✓
Bundle set 3	Form 5 (PR5&PR6a or PR5&PR6b)	100 student booklets			✓	✓
Bundle set 4	Form 201 (PR1)	10 anchor booklets	◆			
Bundle set 5	Form 202 (PR2)	10 anchor booklets	◆			
Bundle set 6	Form 203 (PR3)	10 anchor booklets	◆	◆		
Bundle set 7	Form 204 (PR4)	10 anchor booklets	◆	◆		
Bundle set 8	Form 205 (PR5)	10 anchor booklets		◆		
Bundle set 9	Form 206 (PR6a) or 207 (PR6b)	10 anchor booklets		◆		
Science (trend)	Forms (Clusters)	Number of Booklets per Form	Coder IDs			
			11 (bilingual)	12	13 (bilingual)	14
Bundle set 1	Form 7 (PS1&PS2)	100 student booklets	✓	✓		✓
Bundle set 2	Form 8 (PS3&PS4)	100 student booklets	✓	✓	✓	✓
Bundle set 3	Form 9 (PS5&PS6)	100 student booklets			✓	✓
Bundle set 4	Form 101 (PS1)	10 anchor booklets	◆			
Bundle set 5	Form 102 (PS2)	10 anchor booklets	◆			
Bundle set 6	Form 103 (PS3)	10 anchor booklets	◆	◆		
Bundle set 7	Form 104 (PS4)	10 anchor booklets	◆	◆		
Bundle set 8	Form 105 (PS5)	10 anchor booklets		◆		
Bundle set 9	Form 106 (PS6)	10 anchor booklets		◆		

**Notes:**

“✓” denotes the coder should code 100 student booklets for the specific form as a bundle set. “◆” denotes the coder should code 10 anchor booklets for the specific form as a bundle set.

Paper-based Mathematics, Reading and Science assessments are referred as PM, PR and PS in this table. The number following PM, PR, and PS is the Cluster number. For instance, PM1 represents Cluster 1 in Mathematics domain.

Mathematics and Reading domains have two versions of item cluster 06: 06A and 06B. Each PISA participant selected one or the other version to administer.

CBA participants' coder ID is three-digit; while PBA participants' coder ID is two-digit.

## Within-country and across-country coder reliability

Reliable human coding is critical for ensuring the validity of assessment results within a country, as well as the comparability of assessment results across countries. Coder reliability in PISA 2015 was evaluated and reported at both within- and



across-country levels. The evaluation of coder reliability was made possible by the design of multiple coding - a portion or all of the responses from each human-coded constructed-response item were coded by at least two human coders.

The purpose of evaluating the **within-country coder reliability** was to ensure coding reliability within a country and identify any coding inconsistencies or problems in the scoring process so they could be addressed and resolved earlier in the process. The evaluation of within-country coder reliability was carried out by the multiple coding of a set of student responses, assigning identical student responses to different coders so those responses were coded multiple times within a country. To multiple code all student responses in an international large-scale assessment like PISA is not economical, so a coding design combining multiple coding and single coding was used to reduce national costs and coder burden. In general, a set of 100 responses per human-coded item was randomly selected from actual student responses to be multiple coded. The rest of the student responses needed to be evenly split among coders to be single coded.

Accurate and consistent scoring within a country does not necessarily mean that coders from all countries are applying the coding rubrics in the same manner. Coding bias may be introduced if one country codes a certain response differently than other countries. Therefore, in addition to within-country coder reliability, it was also important to check the consistency of coders across countries. The evaluation of **across-country coder reliability** was made possible by the multiple coding of a set of anchor responses. In each country, two coders in each domain had to be bilingual in English and the language of assessment. These coders were responsible for coding the set of anchor responses in addition to any student responses assigned to them. For each constructed-response item, a set of ten anchor responses in English was provided. These anchor responses were answers obtained from real students and their authoritative coding were not released to the countries. Since countries using the same mode of administration coded the same anchor responses for each human-coded constructed-response item, their coding results on the anchor responses could be compared to each other.

## CODER RELIABILITY STUDIES

Coder reliability studies were conducted to evaluate consistency of coding of human-coded constructed-response items within and across the countries participating in PISA 2015. The studies were based on 59 CBA countries (for a total of 72 country-by-language groups) and 15 PBA countries (for a total of 17 country-by-language groups) with sufficient data to yield reliable results.<sup>4</sup> The coder reliability studies were conducted for three aspects of coder reliability:

- the domain-level proportion agreement
- the item-level proportion agreement
- the coding category distributions of coders on the same item.

Proportion agreement and coding category distribution are the main indicators of coder reliability used in PISA 2015.

- *Proportion agreement* refers to the percentage of each coder's coding that matched the other coders' coding on the identical set of multiple-coded responses for an item. It can vary from 0 (0% agreement) to 1 (100% agreement). Each country was expected to have an average within-country proportion agreement of at least 0.92 (92% agreement) across all items, with a minimum 85% agreement for any one item.
- *Coding category distribution* refers to the aggregation of the distributions of coding categories (such as "full credit", "partial credit" and "no credit") assigned by a coder to two sets of responses: a unique set of 100 responses for multiple coding and responses randomly allocated to the coder for single coding. Notwithstanding that negligible differences of coding categories among coders were tolerated, the coding category distributions between coders were expected to be statistically equivalent based on the standard chi-square distribution due to the random assignment of the single-coded responses.

### Domain-level proportion agreement

The average within-country agreement by domain in PISA 2015 exceeded 92% in each domain across the 89 country-by-language groups with sufficient data (see Tables 13.6 and 13.7). The difference between CBA and PBA participants' average proportion agreements in each of the mathematics, reading and trend science domain was less than 0.5%. Within each mode, the within-country agreements between domains was not significantly different, either. The mathematics domain had higher agreement (97.5% for CBA; 97.5% for PBA) than the other domains. The reading domain also had agreement higher than 95% (95.6% for CBA; 95.8% for PBA). The trend science domain had an average agreement of 94.2% for CBA and 94.7% for PBA. The new science domain for CBA also had an average agreement of 94.2%. The financial literacy domain had slightly lower agreement (93.7% for CBA) than the other domains.

Across-country agreement by domain in PISA 2015 exceeded 92% when averaged over all the 72 CBA country-by-language groups (see Table 13.6). The PBA participants had lower across-country agreement than the CBA participants on average (see Table 13.6 and 13.7). The difference in domain-level proportion agreement between CBA and PBA is 3.3% for mathematics, 3.9% for reading and 5.0% for trend science. Domain-level agreement was the highest in the mathematics domain for both CBA and PBA responses (97.2% for CBA; 94.0% for PBA). For the CBA participants, the reading, trend science, new science and financial literacy domain had across-country agreement at similar levels, ranging between 93.1% and 93.9%. For the PBA participants, the average across-country agreements of the reading and trend science domains were 90.0% and 88.6%, respectively, slightly lower than the criterion but still acceptable.

**[Part 1/2]**  
**Table 13.6 Summary of within-country and across-country agreement (%) per domain for CBA participants**

Computer-based participants (country-by-language unit)	Within-country agreement					Across-country agreement				
	Mathematics (trend)	Reading (trend)	Science (trend)	Science (new)	Financial literacy (trend and new)	Mathematics (trend)	Reading (trend)	Science (trend)	Science (new)	Financial literacy (trend and new)
Australia – English	97.8	92.6	91.7	90.0	93.5	99.4	95.2	99.2	93.3	94.7
Austria – German	97.1	96.3	94.0	95.1		98.1	93.7	96.1	98.0	
Belgium (Flemish) – Dutch	97.5	96.3	93.4	93.5	93.4	97.8	95.9	93.2	94.0	91.6
Belgium (French) – French	98.5	96.8	96.7	96.9		98.9	97.9	97.9	97.0	
Canada – English	96.6	93.6	88.7	89.2	92.2	97.2	95.7	91.8	93.3	95.3
Canada – French	96.9	94.0	89.1	90.3	89.6	96.4	95.6	92.5	93.0	91.6
Chile – Spanish	96.5	94.1	95.1	94.7	92.6	97.9	92.0	94.3	95.7	92.8
Czech Republic – Czech	97.9	96.3	94.2	93.8		98.9	95.1	95.0	95.0	
Denmark – Danish	98.8	97.5	96.6	97.3		98.9	94.1	96.1	93.0	
Estonia – Estonian	95.7	95.4	94.1	93.4		97.8	94.2	94.3	94.3	
Estonia – Russian	95.8	94.1	93.0	92.2		97.8	93.8	94.6	95.2	
Finland – Finnish	99.4	98.2	94.9	94.9		98.6	95.2	95.9	96.0	
France – French	97.8	98.6	95.5	95.0		98.6	94.7	93.9	94.7	
Germany – German	96.5	94.8	93.4	92.3		96.9	94.4	92.9	95.7	
Greece – Greek	96.1	96.2	91.7	92.3		96.9	96.0	93.6	92.3	
Hungary – Hungarian	98.5	94.6	95.6	95.1		96.9	95.2	96.1	96.3	
Iceland – Icelandic	97.7	95.9	95.3	95.0		97.8	96.4	96.1	94.3	
Ireland – English	97.3	94.2	93.7	92.8		97.8	94.5	93.9	94.0	
Israel – Arabic	96.8	96.6	93.8	94.3		90.7	95.5	93.4	89.9	
Israel – Hebrew	96.3	95.3	94.3	94.3		98.3	91.3	95.2	93.2	
Italy – Italian	98.8	94.0	93.2	93.3	93.6	98.5	93.8	92.3	93.9	93.7
Japan – Japanese	97.6	97.4	94.9	96.0		98.1	91.7	92.9	93.0	
Korea – Korean	98.5	97.7	97.0	96.4		98.6	93.3	94.3	90.3	
Latvia – Latvian	95.7	92.5	92.3	94.0		95.3	93.6	93.6	90.7	
Latvia – Russian	96.3	93.4	91.5	92.1		96.1	93.7	93.0	90.7	
Luxembourg – German	97.6	97.1	96.6	97.4		97.8	96.4	96.1	95.3	
Luxembourg – French	98.1	97.3	97.2	97.1		98.3	97.7	96.3	96.2	
Mexico – Spanish	96.7	94.1	92.0	90.5		94.4	93.1	94.3	92.3	
Netherlands – Dutch	99.0	98.7	94.2	95.8	92.2	98.3	96.4	95.4	96.0	94.7
New Zealand – English	97.9	94.4	94.2	93.8		98.3	94.3	95.4	95.8	
Norway – Bokmål	98.0	95.7	96.0	96.4		97.8	95.6	96.1	95.3	
Poland – Polish	98.6	97.3	95.6	94.2	94.5	98.1	94.7	95.0	95.0	94.1
Portugal – Portuguese	97.9	97.5	95.7	95.6		99.4	96.2	95.0	95.0	
Slovak Republic – Slovak	97.5	97.7	95.3	95.4	95.3	98.6	96.6	92.9	92.3	94.7
Slovenia – Slovenian	96.4	96.2	94.5	94.1		98.1	96.0	93.6	95.7	
Spain – Catalan	96.0	95.8	93.6	94.0	96.3	97.2	89.2	93.2	91.0	95.6
Spain – Spanish	96.1	94.1	94.1	94.2	94.0	96.4	88.3	93.2	91.3	90.9
Spain – Basque	93.6	95.2	95.5	92.4	93.3	93.3	90.2	90.7	94.5	92.5
Spain – Galician	97.3	96.6	92.6	94.6		98.1	90.9	93.0	92.3	93.1
Sweden – Swedish	97.6	95.1	94.2	94.2		97.5	95.8	95.9	96.2	
Switzerland – German	97.6	98.0	95.3	95.9		98.1	94.9	94.5	93.2	
Switzerland – French	94.9	95.1	92.7	90.6		98.1	95.6	95.4	92.2	
Switzerland – Italian	96.8	95.9	95.3	94.9		96.9	95.0	96.4	93.7	
Turkey – Turkish	97.7	93.8	94.7	94.1		93.9	89.2	94.6	92.7	
United Kingdom excluding Scotland – English	98.1	95.7	92.7	92.5		98.1	95.6	93.9	94.7	
United Kingdom (Scotland) – English	98.1	96.7	94.9	94.8		97.5	96.5	95.4	93.7	
United States excluding Puerto Rico – English	97.3	94.0	91.1	89.4	92.7	99.1	96.3	93.4	93.3	95.6
Mean – OECD	97.3	95.7	94.1	94.0	93.3	97.5	94.4	94.5	93.9	93.6
Median – OECD	97.5	95.8	94.2	94.2	93.4	97.9	95.0	94.3	94.0	93.9



## [Part 2/2]

Table 13.6 Summary of within-country and across-country agreement (%) per domain for CBA participants

	Computer-based participants (country-by-language unit)	Within-country agreement					Across-country agreement				
		Mathematics (trend)	Reading (trend)	Science (trend)	Science (New)	Financial literacy (trend and new)	Mathematics (trend)	Reading (trend)	Science (trend)	Science (New)	Financial literacy (trend and new)
OECD Partners	Brazil – Portuguese	97.2	93.9	92.1	92.4	93.0	86.2	90.7	85.7	79.7	86.3
	Bulgaria – Bulgarian	93.2	87.1	90.7	90.8		95.0	82.9	92.1	91.3	
	B-S-J-G (China)* – Chinese	97.4	96.8	93.1	93.9	94.4	96.9	95.8	93.6	93.7	90.6
	Colombia – Spanish	99.9	98.8	99.5	98.8		98.2	93.5	88.2	88.7	
	Costa Rica – Spanish	97.6	95.3	93.3	93.3		95.6	94.1	82.9	82.0	
	Croatia – Croatian	98.5	95.7	96.2	97.1		98.9	95.1	94.6	94.3	
	Cyprus <sup>1</sup> – Greek	98.5	96.4	93.4	93.8		98.8	95.1	95.0	97.0	
	Dominican Republic – Spanish	97.0	95.9	95.9	96.4		92.4	81.3	96.8	95.0	
	Hong Kong – Chinese	98.0	95.7	95.6	94.4		98.8	94.8	95.7	95.3	
	Lithuania – Lithuanian	98.0	96.7	96.5	96.5	95.7	98.6	95.1	95.0	96.3	94.4
	Macao – Chinese	99.3	96.2	94.6	94.0		99.2	94.7	93.2	93.0	
	Malaysia – English	97.9	95.3	95.6	95.5		98.6	92.8	90.5	91.2	
	Malaysia – Malay	98.4	95.6	94.7	97.1		98.5	95.7	87.4	91.4	
	Montenegro – Serb (Yekavian)	98.9	96.7	94.2	94.8		97.5	93.7	85.6	85.4	
	Peru – Spanish	99.2	96.9	96.2	96.7	96.6	97.6	93.6	93.2	95.0	95.3
	Qatar – Arabic	98.9	94.7	93.5	93.9		95.3	92.3	93.1	88.7	
	Qatar – English	97.4	94.8	92.7	93.4		97.2	94.4	88.9	91.0	
	Russian Federation – Russian	98.1	95.8	92.7	92.9	94.5	97.5	97.2	92.9	95.7	94.4
	Singapore – English	98.2	95.5	95.5	94.8		96.9	95.9	95.0	94.7	
	Chinese Taipei – Chinese	97.3	96.3	96.2	95.8		99.4	95.1	95.0	96.7	
	Thailand – Thai	98.3	97.3	95.5	96.5		98.9	95.3	95.7	95.7	
	Tunisia – Arabic	99.5	97.0	95.5	95.2		95.3	90.0	87.9	86.7	
	United Arab Emirates – Arabic	97.8	94.1	90.5	91.7		94.1	88.9	92.1	88.0	
	United Arab Emirates – English	96.4	93.5	92.7	92.5		96.2	94.6	92.9	92.0	
	Uruguay – Spanish	97.5	93.8	95.3	94.1		97.1	92.3	92.1	93.3	
Mean – Partners		97.9	95.4	94.5	94.6	94.8	96.8	93.0	91.8	91.7	92.2
Median – Partners		98.0	95.7	94.7	94.4	94.5	97.5	94.4	92.9	93.0	94.4
Mean – All		97.5	95.6	94.2	94.2	93.7	97.2	93.9	93.6	93.1	93.3
Median – All		97.6	95.8	94.3	94.2	93.6	97.8	94.7	93.9	93.7	94.1

\* B-S-J-G (China) refers to the four PISA-participating Chinese provinces: Beijing, Shanghai, Jiangsu and Guangdong.

1. Note by Turkey: The information in this table with reference to « Cyprus » relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognizes the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the “Cyprus issue”.

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognized by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Note: PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

## Item-level proportion agreement

In terms of student responses, all CBA participants had only five or fewer items with proportion agreement lower than 85% in mathematics, new science and financial literacy (see Table 13.8). 96% of them had proportion agreement higher than 85% for every item in those three domains. More than 97% of CBA participants had five or fewer items with proportion agreement lower than 85% in the reading and trend science domains. In terms of student responses, 94% of PBA participants had only five or fewer items with proportion agreement lower than 85% in mathematics; 83% did in reading and trend science.

**Table 13.7 Summary of within-country and across-country agreement (%) per domain for PBA participants**

	Paper-based participants (country-by-language unit)	Within-country agreement			Across-country agreement		
		Mathematics (trend)	Reading (trend)	Science (trend)	Mathematics (trend)	Reading (trend)	Science (trend)
Members	United States (Puerto Rico) <sup>1</sup> – Spanish	98.0	94.8	95.5	95.8	94.4	95.6
Partners	Albania – Albanian	97.5	95.9	96.4	91.7	87.6	86.3
	Algeria – Arabic	85.8	81.9	78.3	80.9	85.6	84.7
	Argentina – Spanish	99.5	98.5	95.0	96.8	93.5	95.0
	Georgia – Georgian	95.3	95.6	97.8	90.7	90.7	88.1
	FYROM – Macedonian	97.8	98.8	98.9	95.9	91.7	74.5
	FYROM – Albanian	98.1	99.1	99.2	95.9	89.7	79.2
	Indonesia – Indonesian	96.9	96.6	95.5	93.8	92.5	90.2
	Jordan – Arabic	99.6	99.6	98.7	95.3	84.3	90.2
	Kosovo – Albanian	98.2	92.5	87.8	97.0	89.9	89.1
	Lebanon – English	99.3	97.6	98.8	96.5	86.8	93.8
	Lebanon <sup>2</sup> – French	99.5	99.2	98.2	NA	NA	NA
	Malta – English	97.7	94.6	92.3	98.0	94.4	95.0
	Moldova – Romanian	99.2	99.4	98.1	97.2	90.5	95.0
	Romania – Romanian	99.4	97.4	98.2	85.2	87.6	85.6
	Trinidad and Tobago – English	96.2	90.2	87.6	96.1	91.9	89.8
	Viet Nam – Vietnamese	99.3	97.0	94.1	96.4	89.6	85.5
	Mean – Partners	97.5	95.9	94.7	93.8	89.7	88.1
	Median – Partners	98.2	97.2	97.1	95.9	89.9	89.1
	Mean – All	97.5	95.8	94.7	94.0	90.0	88.6
	Median – All	98.1	97.0	96.4	95.9	90.2	89.5

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

2. Lebanon did not produce coded anchor responses in French.

Notes: New science and financial literacy are computer-based assessment domains only in the main survey.

PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

**Table 13.8 Percentages of CBA and PBA participants with a different number of items for which proportion agreement is lower than 85%**

Mode	Number of participants	Number of items with proportion agreements lower than 85 %	Mathematics (trend)	Reading (trend)	Science (trend)	Science (new)	Financial literacy (trend and new)
CBA	72	N = 0	96%	83%	85%	86%	84%
		1 ≤ N ≤ 5	4%	14%	13%	14%	16%
		6 ≤ N ≤ 10	0%	1%	3%	0%	0%
		N > 10	0%	1%	0%	0%	0%
PBA	17	N = 0	76%	59%	59%	NA	
		1 ≤ N ≤ 5	18%	24%	24%		
		6 ≤ N ≤ 10	0%	6%	12%		
		N > 10	6%	12%	6%		

Notes: CBA stands for computer-based assessment and PBA for paper-based assessment.

"Item" in the table refers to "human-coded constructed-response item".

PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

Only 19 out of the 72 CBA participants administered the financial literacy domain.

New science and financial literacy are CBA domains only in the main survey.

The summary in the table is based on student responses rather than anchor responses.

As shown in Table 13.9, not a single item had an international mean lower than 85% over the student responses in both CBA and PBA participants. The overall proportion agreement averaged over each item's international mean was 95% for CBA participants and 96% for PBA participants. Only three items had an international mean lower than 85% over the CBA anchor responses, while the international means of eight items were lower than 85% over the PBA anchor responses. The overall proportion agreement averaged over each CBA item's international mean and each PBA item's international mean was 94% and 91%, respectively.

**Table 13.9 Summary of proportion agreement across the PISA participants**

	Source of response	CBA participants						PBA participants			
		Mathematics (trend)	Reading (trend)	Science (trend)	Science (new)	Financial literacy (trend and new)	Average	Mathematics (trend)	Reading (trend)	Science (trend)	Average
Number of items with average proportion agreement lower than 85% averaged across participants	Student responses	0	0	0	0	0	0	0	0	0	0
	Anchor responses	1	6	4	2	2	3	3	12	9	8
Overall proportion agreement averaged over items' international means	Student responses	97%	95%	94%	94%	94%	95%	97%	96%	95%	96%
	Anchor responses	97%	94%	94%	93%	93%	94%	94%	91%	89%	91%

1. "Item" in the table refers to "human-coded constructed-response item".

Notes: CBA stands for computer-based assessment and PBA for paper-based assessment.

PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

### Coding category distributions

In mathematics, 10% of coders in an average CBA country and 27% of coders in an average PBA country had significantly different coding category distributions from other coders on more than 20% of items (see Table 13.10). In reading, it was 17% for CBA and 52% for PBA, while in trend science, it was 20% for CBA and 66% for PBA. In new science, 35% of coders in an average CBA country had significantly different coding category distributions from other coders on more than 20% of items. In financial literacy, the average was 44%. Although some of those percentages may appear high, all the participants reached an acceptable level of coder reliability which is the minimum of 85% for an item and the average of 92% across all items. For few PBA countries, dissimilar coding category distributions among coders could be occasionally observed along with high proportion agreement on an item. This largely resulted from the different pools of responses upon which coding category distribution and proportion agreement were measured. As mentioned earlier, proportion agreement per item across coders was only based on the unique set of 100 responses for multiple coding; while coding category distribution per item across coders also took into account the randomly assigned responses for single coding. Compared to CBA countries, the randomization of responses was more challenging for PBA countries where the distribution of booklets were handled manually.

**Table 13.10 Percentage of coders whose coding category distributions on more than 20% of coded items were significantly different from other coders, averaged across CBA and PBA participants**

	Mathematics (trend)	Reading (trend)	Science (trend)	Science (new)	Financial literacy (trend and new)
CBA Participants	10%	17%	20%	35%	44%
PBA Participants	27%	52%	66%	NA	NA

Notes: CBA stands for computer-based assessment and PBA for paper-based assessment.

The summary in the table is based on both student responses and anchor responses.

"Item" in the table refers to "human-coded constructed-response item".

New science and financial literacy are CBA domains only in the main survey.

PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

Across all the CBA participants, the percentage of items over which more than two coders' coding category distributions were significantly different from other coders was 6% in mathematics, 14% in reading, 8% in trend science, 13% in new science and 13% in financial literacy (see Table 13.11). Across all the PBA participants, the percentage of items over which more than two coders' coding category distributions were significantly different from other coders was 17% in mathematics, 38% in reading and 23% in trend science (see Table 13.11). Although some of those percentages for PBA participants may appear high, all the participants have reached an acceptable level of coder reliability which is the minimum of 85% for an item and the average of 92% across all items.

**Table 13.11 Percentages of participant × item pairs that have more than two coders' coding category distributions significantly different from other coders**

	Mathematics (trend)	Reading (trend)	Science (trend)	Science (new)	Financial literacy (trend and new)
CBA	6%	14%	8%	13%	13%
PBA	17%	38%	23%	NA	NA

Notes: CBA stands for computer-based assessment and PBA for paper-based assessment.

The summary in the table is based on both student responses and anchor responses.

"Item" in the table refers to "human-coded constructed-response item".

PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities.

The scales on which the PISA statistical framework is built are only as good as the scores used to establish them. In sum, the results from the coder reliability studies revealed that the coding designs that were tailored to meet every PISA participant's specific survey needs and the availability of coders were executed well, especially for CBA human-coded responses. The management of the coding process went smoothly and efficiently, with less involvement from the National Project Managers than necessary in previous cycles. CBA participating countries produced more complete and consistent coding data, while PBA participants showed some errors in the handling of the booklets and less reliable human coding. However, PBA participants still achieved acceptable levels of coder reliability amid the challenge of handling the booklet bundles manually.

### **Notes**

1. PISA participants can be a country, a region, an economy, or a subsample within the former three types of entities. In this chapter, the generic terms "countries" or "participants" are used for the purpose of simplicity.
2. For a better understanding of the PISA coding designs, it is recommended that the descriptions of the PISA assessment designs in Chapter 2 be read first as important background information.
3. In mathematics, there was an additional cluster, as instead of M06 there was M06A and M06B. Since countries could only choose M06A or M06B, but not both, the actual number of clusters in each domain is six rather than seven. The same is true for clusters R06A and R06B in reading.
4. Coding data from Kazakhstan (Kazakh) and Kazakhstan (Russian) were not included in this analysis and all human-coded responses were excluded from the calculation of proficiency estimates.



14

## Data adjudication

<b>Introduction .....</b>	264
<b>General outcomes .....</b>	267



## INTRODUCTION

The PISA Technical Standards (see Annex F) specify the way in which PISA must be implemented in each country, economy, and adjudicated region.<sup>1</sup> International contractors monitor the implementation in each of these and adjudicate on their adherence to the standards. This chapter describes the process used to adjudicate the implementation of PISA 2015 in each of the adjudicated entities (i.e. the participating countries, economies, and adjudicated regions<sup>2</sup>) and gives the outcomes of data adjudication that are mainly based on the following aspects:

- the extent to which each adjudicated entity met PISA sampling standards
- the outcomes of the adaptation, translation, and verification process
- the outcomes of the PISA Quality Monitoring visits
- the quality and completeness of the submitted data
- the outcomes of the international coding review.

## PISA 2015 Technical Standards

The areas covered in the PISA 2015 Technical Standards include the following:

### **Data Standards**

- target population and sampling
- language of testing
- field trial participation
- adaptation of tests, questionnaires, and school-level manuals and scripts
- translation of tests, questionnaires, and school-level manuals and scripts
- test administration
- implementation of national options
- security of the material
- quality monitoring
- printing of material
- response coding
- data submission

### **Management standards**

- communication with the international contractors
- notification of international and national options
- schedule for submission of materials
- drawing samples
- management of data
- archiving of materials

### **National involvement standards**

- national feedback

## Implementing the standards – quality assurance

National Project Managers of participating countries, economies, and adjudicated regions are responsible for implementing the standards based on the international contractors' advice as contained in the various operational manuals and guidelines. Throughout the cycle of activities for each PISA survey, the international contractors carried out quality assurance activities in two steps. The first step was to set up quality control using the operational manuals, as well as the agreement processes for national submissions on various aspects of the project. These processes gave the international contractor staff the opportunity to ensure that PISA implementation was planned in accordance with the PISA 2015 Technical Standards and to provide advice on taking rectifying action when required and before critical errors occurred. The second step was quality monitoring, which involved the systematic collection of data that monitored the implementation of the assessment in relation to the standards. For data adjudication, it was the information collected during both the quality control and quality monitoring activities that was used to determine the level of compliance with the standards.



## Information available for adjudication

The international contractors' quality monitoring of a country's data collection is carried out from a range of perspectives during many stages of the PISA cycle. These perspectives include monitoring a country's adherence to the deadlines, communication from the sampling contractor about each country's sampling plan, information from the language verification team, data from the PISA Quality Monitors, and information gathered from direct interviews at National Project Manager and Coder Training meetings. The information was combined together in the database so that:

- indications of non-compliance with the standards could be identified early on in order to enable rectifying measures
- the point at which the problem occurred could be easily identified
- information relating to the same PISA standard could be cross-checked between different areas or sources.

Many of these data collection procedures refer to specific key documents, specified in the National Project Manager's Manual and the Sampling Manual in particular. These are procedures that the international contractors require for field trial and main survey preparation from each National Centre. The data adjudication process provides a motivation for collating and summarising the specific information relating to PISA standards collected in these documents, combined with information collected from specific quality monitoring procedures such as the PISA Quality Monitor visits and from information in the submitted data.

The quality monitoring information was collected from various quality monitoring instruments and procedures and covered the following main administrative areas:

- international contractors' administration and management: information relating to administration processes, agreement of adaptation spreadsheets, submission of information
- data analysis: information from item level reports, from the field trial data, and from data cleaning steps, including consistency checks
- school-level materials: information from the agreement of adaptations to test administration procedures and field operations
- Final Optical Check team: information from the pre- and post-main survey Final Optical Checks of main survey booklets
- main survey review: information provided by the National Project Managers in the main survey Review Questionnaire
- National Centre quality monitoring: information gathered through interviews conducted during meetings of National Project Managers or at other times
- co-ordination of PISA Quality Monitor activities including recruitment
- PISA Quality Monitor country reports: information gathered via the Data Collection Forms from PISA Quality Monitors and through their interactions with School Co-ordinators and Test Administrators
- sampling: information from the submitted data such as school and student response rates, exclusion rates and eligibility problems
- translation: information relating to the verification and translation process
- National Centre Test Administrator or School Associate trainings
- National quality monitoring issues
- data cleaners: issues identified during the data cleaning checks and from data cleaners' reports
- item developers: issues identified in the coder query service and training of coders;
- data processing: issues relating to the eligibility of students tested
- questionnaire data: issues relating to the questionnaire data in the national questionnaire reports provided by the international contractor
- questionnaire Final Optical Check: issues arising from the Final Optical Check of the questionnaires.

## Quality monitoring reports

There were two types of PISA quality monitoring reports: The Session Report Form containing data for each session in each school, and the Data Collection Form detailing the general observations across all schools visited by PQMs. The Session Report Form was completed by the Test Administrator after each test session and also contained data related to test administration. The data from this report were recorded by the National Centre and submitted as part of the national dataset to ETS. The PISA Quality Monitor reports contained data related to test administration in selected schools, and the PISA quality monitoring data were collected independently of the National Project Manager. Additional information on all the standards was also noted in the main survey Review. The main survey Review was self-declared by the National Project Manager.

## Data adjudication process

The main aim of the adjudication process is to make a judgement on each national dataset in a manner that is transparent, based on evidence, and defensible. The data adjudication process achieved this through the following steps:

**Step 1:** Quality control and quality monitoring data were collected throughout the survey administration period.

**Step 2:** Data collected from both quality control and quality monitoring activities were entered into a single quality assurance database.

**Step 3:** Experts compiled country-by-country reports that contained quality assurance data for key areas of project implementation.

**Step 4:** Experts considered the quality assurance data that were collected from both the quality control and quality monitoring activities to make a judgement. In this phase, the experts collaborated with the international contractors to address any identified areas of concern. Where necessary, the relevant National Project Manager was contacted through the contractors. At the end of this phase experts constructed, for each adjudicated dataset, a summary detailing how the PISA Technical Standards had been met.

**Step 5:** The adjudication group, formed by representatives of the OECD and of international contractors, the Technical Advisory Group and the Sampling Referee, reviewed the reports and made a determination with regard to the quality of the data from each adjudicated entity.

Monitoring compliance to any single standard occurred through responses to one or more quality assurance questions regarding test implementation and national procedures which may come from more than one area. For example, the session report data were used in conjunction with the PISA Quality Monitor reports, computer system tracking of timings, and information from the adaptation of national manuals to assess compliance with the PISA session timing standard (Standard 6.1, Annex F).

Information was collected in relation to these standards through a variety of mechanisms:

- through PISA Quality Monitor reports
- through the field trial and main survey reviews
- through information negotiated and stored on the PISA Portal website (the portal which was used in PISA 2015)
- through a system database specific to the implementation of PISA tasks
- through the formal and informal exchanges between the international contractors and National Centres over matters such as sampling, translation and verification, specially requested analyses (such as non-response bias analysis)
- through a detailed post-hoc inspection of all main survey assessment materials (test booklets)
- through the data cleaning and data submission process.

For PISA 2015, an adjudication database was developed to capture, summarise, and store the most important information derived from these various information sources. The staff members of the international contractor who led each area of work were responsible for identifying relevant information and entering it into the database. This means that at the time of data adjudication, relevant information was easily accessible for making recommendations about the fitness of use of data from each PISA adjudicated entity.

The adjudication database captured information related to the major phases of the data operation: field operations, sampling, computer-based problem solving and computer-based assessment of Financial Literacy (where applicable), questionnaires, and tests. Within each of these phases, the specific activities are identified, and linked directly to the corresponding standards.

Within each section of the database, specific comments are entered that describe the situation of concern, the source of the evidence about that situation, and the recommended action. Each entry is classified as serious, minor, or of no importance for adjudication. Typically, events classified as serious would warrant close expert scrutiny and possibly action affecting adjudication outcomes. For example, cognitive data for Kazakhstan were found to be inconsistent across human-coded and machine-scored items, and upon further investigation, the coding of human coded items was invalidated, resulting in the exclusion of Kazakhstan's performance scores from international comparisons and comparisons with results for Kazakhstan from previous years. Events classified as minor would typically not directly affect adjudication outcomes but will be reported back to National Centres to assist them in reviewing their national procedures.



## Data adjudication

It was expected that the data adjudication would result in a range of possible recommendations to the PISA Governing Board. Some possible, foreseen recommendations included:

- that the data be declared fit for use
- that some data be removed for a particular country, economy, or adjudicated region, such as the removal of data for some open-ended items or the removal of data for some schools
- that rectifying action be performed by the National Project Manager, such as providing additional evidence to demonstrate that there was no non-response bias, or rescored open-ended items
- that the data not be endorsed for use in certain types of analyses
- that the data not be endorsed for inclusion in the PISA 2015 database.

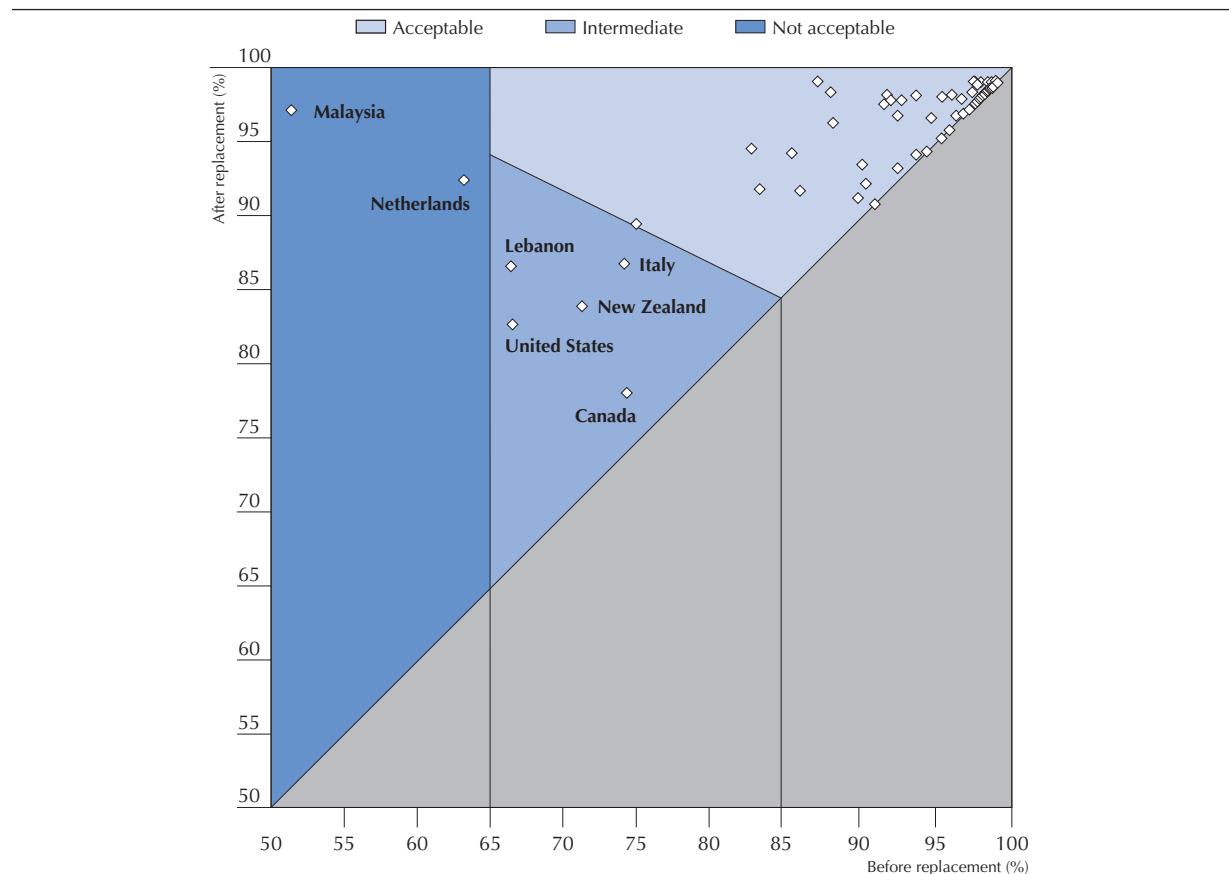
Throughout PISA 2015, the international contractors concentrated their quality control activities to ensure that the highest scientific standards were met. However, during data adjudication a wider definition of quality was used, especially when considering data that were at risk. In particular, the underlying criterion used in adjudication was fitness for use; that is, data were endorsed for use if they were deemed to be fit for meeting the major intended purposes of PISA.

## GENERAL OUTCOMES

### Overview of response rate issues

The PISA school response rate requirements are discussed in Chapter 4. Figure 14.1 is a scatter plot of the attained PISA school response rates before and after replacements. Those countries that are plotted in the light blue shaded region were regarded as fully satisfying the PISA school response rate criterion.

■ Figure 14.1 ■  
Attained school response rates



Source: PISA 2015 Technical Report, Tables 11.3 and 11.5.

Seven countries – Canada, Italy, Lebanon, Malaysia, the Netherlands, New Zealand and the United States – failed to meet the school non-response rate (see Figure 14.1). After reviewing the sampling outcomes, the consortium asked these seven countries to provide additional data that would assist the consortium in making a balanced judgement about the threat of the non-response to the accuracy of inferences which could be made from the PISA data.

Belgium, Hong Kong (China), the United Kingdom and, among adjudicated subnational entities, the Flemish community (Belgium) and Massachusetts (United States), had a response rate below the 85% level before the use of replacement schools but cleared the acceptable level after the replacement schools were included.

One country – Trinidad and Tobago – fell short of the student response-rate standard; however, in consideration of the fact that there were no student-level exclusions and that the achieved response rate (79.4%) was very close to the acceptable rate of 80%, the international contractor determined that the data were acceptable (see Table 11.7).

### **Detailed country comments**

It is important to recognise that PISA data adjudication is a late but not necessarily final step in the quality assurance process. By the time each country was adjudicated at the Technical Advisory Group meeting in June 2016, the quality assurance and monitoring processes outlined earlier in this chapter and in Chapter 7 had been implemented. Data adjudication focused on residual issues that remained after these quality assurance processes had been carried out.

The remaining issues fall under two broad categories: (1) adaptations to the recommended international standard procedures in a country's data collection plan, and (2) a failure to meet international standards at the implementation stage.

#### ***Departures from standard procedures in the national data collection plan***

With such a broad and diverse range of participation, it is to be expected that the international best practice approaches to data collection articulated in the PISA Technical Standards document may not be achieved in all national and local contexts. This may be the case for a number of reasons. For example, it may be contrary to national protocols to have unannounced visits of quality monitors to schools to observe test administration. Or, it may not be possible for teachers from very remote or very small schools to leave their schools to attend training in the mechanics of PISA test administration. Typically these were discussed with international contractor experts in advance of the assessment, and alternative approaches were considered jointly between the National Project Manager and the international contractor. In isolated departures from best practice in cases such as these, a judgement might easily be made by international contractor experts that there was minimal risk to the quality of the data collection plan. Such isolated departures are not reported in the country summaries below.

On the other hand, it may not have been straightforward to determine in advance of the assessment how more extensive or multiple departures from PISA Technical Standards may interact with each other and with other aspects of a country's data collection plan. Cases such as these were considered as part of the data adjudication process and are included in the country summaries below.

#### ***Departures from standards arising from implementation***

Departures from the standards at the implementation stage range from errors within the National Centre (e.g., during the final stages of preparing materials, or in the administration of the coding operation following data collection), a failure to meet documented targets during data collection, for example a shortfall from the minimum school and student sample sizes.

A component of the data adjudication process was to consider the cases of multiple, or more complex departures from the PISA standard procedures, as well as to consider the impact of errors or shortfalls across all aspects of each country's data collection plan and implementation, and make an evaluation with respect to the quality and international comparability of the PISA results. Notable departures from the standards are reported in the country summaries below.

Several countries exceeded the limit on student- and school-level exclusions (5% at most; see Tables 11.1 and 11.2). In countries where other violations of sampling standards were observed or where the combined level of exclusions exceeded 6%, further information was requested to support the case that no bias would result from exclusions. The number of such cases shows a notable increase over the level observed in previous cycles; it is unclear whether the increase in exclusion rates observed in several countries over previous cycles must be attributed to the first implementation of computer-based assessment, to an increase in migrant populations, or to local, idiosyncratic factors.



A small number of countries failed to reach the required minimum sample sizes of 5 400 students and 150 schools (4 500 students and 150 schools for countries that tested in paper mode or did not participate in the testing of collaborative problem solving; the numbers for additional adjudicated entities are 1 800 students and 50 schools and 1 500 students and 50 schools, respectively). Such cases were considered as part of the data adjudication process. Even a minor deviation in sample size might be considered a substantive enough issue to report, for example in countries where standard errors tend to be higher for a given sample size. On the other hand, minor deviations from these minimal sample sizes (i.e. shortfalls of fewer than 50 students or 5 schools, and in countries that nevertheless achieved comparable standard errors on the major survey estimates) are not reported below.

Particular attention has been paid to the achievement of the specified response rates of 85% for schools, 80% for students within schools and no more than 5% of students excluded from the assessment. Seven countries were required to provide additional data to support the case that no bias would result from failure to meet the response-rate standards.

Anomalies in submitted data, particularly inconsistencies and deviations from the expected patterns, were also investigated; most cases could be explained and solved through a resubmission of data. The two cases that could not be solved are noted below.

If a country is not listed below then it fully met the PISA standards. Further, in the case of minor deviations from the standards, unless otherwise noted, additional data were usually available to suggest the data were suitable for use.

## **Albania**

Analysis of the data submitted by Albania suggested that the PISA Technical Standards were not fully met. Indeed, the relationships between student achievement and student background characteristics collected through self-report questionnaires were, without exceptions, very weak, deviating from associations found in Albania in previous cycles and from the patterns observed in other countries. There was no association, for instance, between reading performance and the reported number of books at home. However, school-level associations and relationships with student characteristics collected with student tracking forms (such as gender or grade) conformed with expectations. A mismatch between cognitive booklets and questionnaire booklets, whereby the same student identifier was used for different students within the same school, is suspected.

The PISA 2015 international database therefore does not include information collected through student questionnaires for Albania. An additional dataset, which uses different student identifiers, contains this information. No attempt should be made to link the student data included in the international PISA database with the additional dataset for Albania.

## **Argentina**

In Argentina the review of sampling outcomes revealed that the national defined target population (the population of all 15-year-old students enrolled in grade 7 and above in schools listed in the sampling frame for PISA, and not marked for exclusion) deviated significantly from the desired target population.

The sampling frame submitted in February 2015 contained much fewer schools than in previous cycles, but showed good agreement between the sum of school-level enrolment figures and the overall enrolment expected, given the experience of previous cycles and national enrolment statistics. Despite reassurance about the completeness of the sampling frame, the actual enrolment figures collected from sampled schools in June-July 2015 fell significantly short (by about 30%) of the expected number of 15-year-old students in each school. As a result, the coverage of the sampling frame enrolment (Coverage index 4), in 2015, was only 69%. Further information was therefore requested to exclude that the "missing" students were enrolled in schools that were not listed on the sampling frame.

Upon further investigation, it was found that three categories of schools were omitted from the sampling frame:

- Schools that had been created or renamed between 2013 and 2015. Due to an ongoing restructuring policy that merged lower- and upper-secondary schools into new unified secondary schools, the number of students attending newly created or renamed schools was higher than usual.
- Schools that were listed incorrectly in the official registry of educational institutions "administrative headquarters" with no students. This error in the official registry of educational institutions, one of the sources used to identify schools with 15-year-old students, was detected only in October 2015.
- Finally, some provincial authorities did not include all rural schools as they had very few students.



Together, these omissions resulted in the exclusion of well over 10% of the desired target population from the sampling frame, creating a significant threat of non-coverage bias. However, the investigation conducted in June 2016 also highlighted that no eligible school in the adjudicated entity of Ciudad Autónoma de Buenos Aires was missing from the original sampling frame.

The failure to meet standard 1.7 about the definition of the target population and school exclusions implies that Argentina's results may not be representative of the whole country, and may therefore not be comparable to those of other countries or to results for Argentina from previous years. Data for Argentina were therefore not included in the international dataset, and are available as a separate dataset. Data for Ciudad Autónoma de Buenos Aires, on the other hand, were deemed to be acceptable and are fully included in the international dataset.

## **Australia**

There was a total of 5.31% exclusions in Australia; data were included in the final database.

## **Canada**

There was a total of 7.49% exclusions in Canada. Exclusion rates however were only marginally higher than in previous cycles, and, as in 2012, exclusions were mostly due to students with special needs.

Canada had a weighted school response rate of 74.5% before replacement. After replacement, the response rate was 78.6% (a response rate of 90.3% was required in order to fully meet the standard). Further information was sought from Canada to support the case that no notable bias would result from non-responses.

Additional analyses indicated that much of Canada's non-responses came from the relatively large provinces of Alberta, Ontario and Quebec, with the latter province achieving the lowest school response rate among all provinces; all other provinces had acceptable response rates. Consequently, the non-response bias analysis compared the characteristics of the target population and of responding, non-responding, and replacement schools in these three provinces only. In addition to school type and language, assessment results in local assessments were available for all students in Alberta, and at the school level in Ontario (for sampled schools only) and Quebec. Canada presented evidence to show that the characteristics of non-responding schools in Alberta and Ontario were not markedly different from those of respondent schools, while for Quebec the comparison of mean achievement scores at the school level in an assessment of reading and science showed more significant differences.

The adjudication group concluded that no notable bias would result from non-response in the Canadian data, when analysed at the national level, and inclusion in the full range of PISA 2015 reports was recommended. However, caution was invited when reporting data for the province of Quebec in isolation, due to a possible non-response bias.

## **Denmark**

There was a total of 5.04% exclusions in Denmark; data were included in the final database.

## **Estonia**

There was a total of 5.52% exclusions in Estonia; data were included in the final database.

## **Italy**

Italy had a weighted school response rate of 74.4% before replacement. After replacement, the response rate was 87.5% (a response rate of 90.3% was required in order to fully meet the standard).

Additional analysis indicated that a number of characteristics were balanced across respondents and non-responding schools. The higher level of refusals to participate compared to past administrations was explained by the burden created by bigger-than-usual samples, due to the addition of a grade-based sample as national option, in the context of a computer-based administration.

The adjudication group concluded that no notable bias would result from non-response. The data for Italy, therefore, were included in the international database.



## Kazakhstan

During the consistency checks performed on cognitive data prior to scaling, it was found that scores for human-scored items submitted by Kazakhstan were inconsistent with the success rates observed in prior PISA cycles and were almost unrelated to scores on machine-scored (multiple choice) items. Further information was sought to ensure that the coded responses reflected student responses in an authentic way.

Kazakhstan was asked to send 300 randomly selected booklets, in both national languages, for re-coding. This independent coding performed by international coders indicated significant leniency among national coders. However, even the lower scores assigned by international coders deviated from the expected patterns of association with machine-scored items for the same students. The evidence was deemed sufficient to conclude that the data submitted by Kazakhstan did not meet coding standards.

The adjudication group recommended to invalidate all human-scored items. Furthermore, because an assessment that is limited to multiple-choice items would not provide adequate coverage of the constructs measured in PISA and could not be considered comparable to other countries or to past results, the remaining data were not deemed to be fit for inclusion in the international database, and only limited reporting was recommended. Data for Kazakhstan are available as a separate dataset and caution must be exerted when comparing the results for Kazakhstan to past cycles or to results for other countries.

## Latvia

There was a total of 5.07% exclusions in Latvia and there were fewer than the 5400 students specified in the standards for a country or economy participating in the assessment of collaborative problem solving (4845); data were included in the final database.

## Lebanon

Lebanon had a weighted school response rate of 66.6% before replacement. After replacement, the response rate was 87.3% (a response rate of 94.2% was required in order to fully meet the standard). Additional analysis showed that language of instruction and the distribution of school enrolment did not differ significantly between the original sample and the final responding sample using replacements; no other characteristics were available for a non-response bias analysis. The adjudication group nevertheless concluded that it was unlikely that notable bias would result from non-response in the final database. The data for Lebanon, therefore, were included in the international database.

## Lithuania

There was a total of 5.12% exclusions in Lithuania; data were included in the final database.

## Luxembourg

There was a total of 8.16% exclusions in Luxembourg. In consideration of the consistency in the level and nature of exclusions across cycles in Luxembourg, data were deemed to be acceptable and included in the database.

## Malaysia

In Malaysia, the weighted response rate among the initially sampled schools (51.4%) fell short of the 65% minimal threshold, and corresponds to an unacceptable response rate (after the use of replacement schools, the response rate was 98.1%). Malaysia submitted a non-response bias analysis which showed that responding replacement schools had significantly better result, on a national examination, than non-responding schools in the original sample. The adjudication group concluded that in this case, non-response may have introduced bias in comparisons of Malaysia's results with those of other countries or with previous years, and recommended limited reporting. Data for Malaysia are included in a separate database.

## Montenegro

There was a total of 5.17% exclusions in Montenegro; data were included in the final database.

## **Netherlands**

In the Netherlands, the weighted response rate among the initially sampled schools (63.3%) fell slightly short of the 65% minimal threshold, and corresponds to an unacceptable response rate (after the use of replacement schools, the response rate was 93.2%). Additional analysis however, using school results on a central examination of mathematics and science subjects, indicated that within each stratum the mean and distribution of results of responding schools (including replacements) did not differ significantly from the target population. The adjudication group concluded that no notable bias would result from non-response in the final database. The data for the Netherlands, therefore, were included in the international database.

## **New Zealand**

There was a total of 6.54% exclusions in New Zealand. Available information indicated that the extra students excluded were all students with limited language ability, recently arrived in the country.

There were fewer than the 5400 students specified in the standards for a country or economy participating in the assessment of collaborative problem solving (4520).

New Zealand also had a weighted school response rate of 71.4% before replacement. After replacement, the response rate was 84.5% (a response rate of 91.8% was required in order to fully meet the standard). Additional analysis, using in particular school results on the New Zealand National Certificate in Educational Achievement, indicated that the mean and distribution of results did not differ significantly between responding schools (including and excluding replacements) and the original sample. The adjudication group concluded that no notable bias would result from non-response in the final database. The data for New Zealand, therefore, were included in the international database.

## **Norway**

There was a total of 6.75% exclusions in Norway. Most of the excess students excluded were students with limited language ability, recently arrived in the country, and data were therefore deemed to be acceptable and included in the final database.

## **Spain**

### **Castile and Leon**

There was a total of 5.08% exclusions in Castile and Leon; data were included in the final database.

### **Catalonia**

There was a total of 5.41% exclusions in Catalonia; data were included in the final database.

### **Valencia**

There was a total of 7.41% exclusions in Valencia; and there were fewer than the 1 800 students specified in the standards for an adjudicated entity participating in the assessment of collaborative problem solving (1611). Data were included in the final database.

## **Sweden**

There was a total of 5.71% exclusions in Sweden. Available information indicated that the extra students excluded were all students with limited language ability, recently arrived in the country, and data were therefore deemed to be acceptable and included in the final database.

## **Trinidad and Tobago**

Trinidad and Tobago had a weighted student response rate of 79.4%, slightly below the standard of 80%. In consideration of the fact that there were no school or student-level exclusions, the adjudication group concluded that no notable bias would result from non-response in the final database. The data for Trinidad and Tobago, therefore, were included in the international database.



## **United Kingdom**

There was a total of 8.22% exclusions in the United Kingdom, and a marked increase in exclusions due to special needs over previous cycles. The national centre for the United Kingdom (excluding Scotland) explained this as a possible unintended consequence of changes in the timing at which information about special needs was collected (in student tracking forms rather than in student lists collected prior to sampling); this could have led school coordinators to mark more students for exclusion. In consideration of the fact that appropriate actions had been taken to limit exclusions once this issue had been detected, data were deemed to be acceptable and included in the final database.

### **Scotland**

There was a total of 6.52% exclusions in Scotland; data were included in the final database.

## **United States**

The United States had a weighted school response rate of 66.7% before replacement. After replacement, the response rate was 83.3% (a response rate of 94.2% was required in order to fully meet the standard). Additional analysis, using data from two school surveys and from the sampling frame, indicated that among originally sampled schools, region and student race/ethnicity differed across responding and non-responding schools. However, after replacement schools were added to the respondents, all available characteristics (school enrolment, control and location, region, race/ethnicity and, for public schools, poverty) were balanced with those of the initially selected school sample, therefore showing that the use of replacement schools substantially reduced the potential for bias. The adjudication group concluded that no notable bias would result from non-response in the final database. The data for the United States, therefore, were included in the international database.

### **Massachusetts (Public schools)**

There were fewer than the 1 800 students specified in the standards for an additional adjudicated entity participating in the assessment of collaborative problem solving (1391). Data were included in the final database.

### **Puerto Rico<sup>3</sup>**

There were fewer than the 1 500 students specified in the standards for an additional adjudicated entity testing in paper mode (1398). Data were included in the final database.

## **Notes**

1. For the remainder of this chapter, we will use the term “country” when referring to a country, economy, or adjudicated region.
2. Not all regions opt to undergo the full adjudication that would allow their results to be compared statistically to all other participating economies and adjudicated regions. For example, the states of Australia are not adjudicated regions, whereas the Flemish Community of Belgium is an adjudicated region.
3. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.





15

## Proficiency scale construction

<b>Introduction .....</b>	276
<b>Development of the described scales .....</b>	277
<b>Defining the proficiency levels .....</b>	279
<b>Reporting the results for pisa science .....</b>	281



## INTRODUCTION

This chapter discusses the methodology used to develop the PISA reporting scales, which describe levels of proficiency in the different PISA domains, and presents the outcomes of the development process for science literacy, the major domain in PISA 2015.

The reporting scales are called “proficiency scales” rather than “performance scales” because they describe what students *typically* know and can do at given levels of proficiency, rather than how individuals who were tested *actually* performed on a single test administration. This emphasis reflects the primary goal of PISA, which is to report general population-level results rather than the results for individual students. PISA uses samples of students and items to make estimates about populations. A sample of 15-year-old students is selected to represent all 15-year-olds in a country and a sample of test items from a large pool is administered to each student. Results are then analysed using statistical models that estimate the likely proficiency of the population, based on this sampling.

The PISA test design makes it necessary to use techniques of modern item response modelling (see Chapter 9) to both estimate the ability of all students taking the PISA assessment and the statistical characteristics of all PISA items.

The PISA data are collected using a rotated test design in which students take different but overlapping tasks. The mathematical model employed to analyse the PISA data is implemented through test analysis software that uses iterative procedures to simultaneously estimate the distribution of students along the proficiency dimension assessed by the test, as well as a mathematical function that describes the association of student proficiency and the likelihood of a correct response for each item on the test. The result of these procedures is a set of item parameters that represents, among other things, locations on a proficiency continuum reflecting the domain being assessed. On that continuum, it is possible to estimate the distribution of groups of students, and thereby the average (location) and range (variability) of their skills and knowledge in this domain. This continuum represents the overall PISA scale in the relevant test domain, such as reading, mathematics, or science.

PISA assesses students and uses the outcomes of that assessment to produce estimates of students’ proficiency in relation to the skills and knowledge being assessed in each domain. The skills and knowledge of interest, as well as the kinds of tasks that represent those abilities, are described in the PISA frameworks (OECD, 2017). For each domain, one or more scales are defined, each ranging from very low levels of proficiency to very high levels. Students whose ability estimate places them at a certain point on a PISA proficiency scale would be more likely to be able to successfully complete tasks at or below that point. Those students would be increasingly *more likely* to complete tasks located at progressively lower points on the scale, and increasingly *less likely* to complete tasks located at progressively higher points on the scale. Figure 15.1 depicts a simplified hypothetical proficiency scale, ranging from relatively low levels of proficiency at the bottom of the figure, to relatively high levels towards the top. Six items of varying difficulty are placed along the scale, as are three students of varying ability. The relationship between the students and items at various levels is described in the figure.

In addition to defining the numerical range of the proficiency scale, it is also possible to define the scale by describing the competencies typical of students at particular points along the scale. The distribution of students along this proficiency scale is estimated, and locations of students can be derived from this distribution and their responses on the test. Those location estimates are then aggregated in various ways to generate and report useful information about the proficiency levels of 15-year-old students within and among participating countries.

The development of a method for describing proficiency in PISA reading, mathematical and scientific literacy occurred in the lead-up to the reporting of outcomes of the PISA 2000 survey and was revised in the lead-up to the PISA 2003, 2006, 2009 and 2012 surveys. Although essentially the same methodology has again been used to develop proficiency descriptions for PISA 2015, a more general statistical model compared to previous cycles was used in the scaling procedure (see Chapter 9 for details).

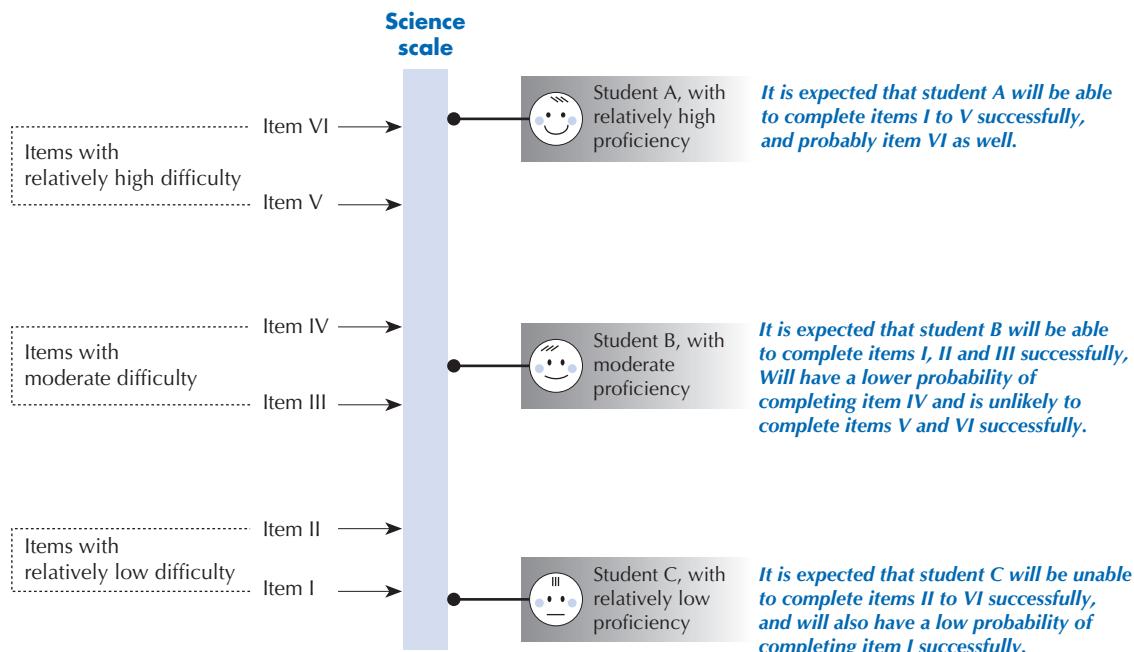
The proficiency descriptions that had been developed for the mathematics domain in PISA 2012, for reading in 2009 and for financial literacy in 2012 were used again to report the 2015 results.

Reporting for science, the major domain in 2015, was linked back to the 2006 proficiency scale and was based on the detailed proficiency level descriptions developed in 2006, the last cycle in which science was the major domain. These proficiency level descriptors were reviewed and revised based on the 2015 data in order to incorporate the new science framework developed for this cycle and the performance of the new computer-based items, including the interactive simulation tasks.



■ Figure 15.1 ■

### Simplified relationship between items and students on a proficiency scale



The science expert group worked with the PISA international contractor to review and revise the sets of described proficiency scales and subscales for PISA science. Similarly, the international contractor worked with the collaborative problem solving expert group to develop the described proficiency scale for that domain.

### DEVELOPMENT OF THE DESCRIBED SCALES

The development of described proficiency scales for PISA has been carried out through a process that typically involves a number of tasks conducted by the expert groups and the item development team. The process of developing the described scales involved several iterations as the data were collected and analysed during the 2015 cycle. It should be noted that, as each PISA cycle builds upon the work implemented in previous cycles, the same tasks are not completed for every domain in every cycle. The following description of the development process focuses on the development of described proficiency scales for science and collaborative problem solving.

### Classification of items

As part of new item development for science and collaborative problem solving, test developers classified all items based on the specifications provided in the framework for each domain. Item classifications for the trend science items were also revised to reflect the 2015 framework. All classifications were reviewed by each of the expert groups and revised as needed.

### Defining the overall proficiency scale

As part of its work in developing the framework for science, the expert group drafted initial descriptors of the levels of scientific literacy, based on the knowledge and competency dimensions defined therein. These descriptors, presented as an initial hypothesis, were shared as part of the framework to allow item developers to design items representing the increase in skills and ability reflected across the levels.

Final item parameters were estimated for the trend and new science items based on analysis of the Main Survey data. Using this information on item performance, the science expert group met over several days and reviewed each of the items and discussed key characteristics that differentiated performance along the proficiency scale. As part of that review process, the initial draft descriptors for each level in the overall proficiency scale were refined and finalised.



Defining the proficiency scale for collaborative problem solving was more challenging because the domain was newly developed for the 2015 cycle. The experts defined a matrix of collaborative problem solving skills in the collaborative problem solving framework that served as the basis for describing performance along the scale. They also set cut-off points along the scale that defined each level of performance. Description and definition of the proficiency scale for collaborative problem solving is also provided in the PISA 2015 Frameworks report (OECD, 2017).

### **Identifying possible subscales**

For each domain in PISA, reporting includes an overall proficiency scale based on the combined results for all items within that domain. In addition, the framework may support subscales based on the various dimensions of the framework. Where subscales are included, they must arise clearly from the domain framework, be meaningful and potentially useful for feedback and reporting purposes, and be defensible with respect to their measurement properties. Thus, the first stage in the process involves having the experts articulate possible reporting subscales based on the most recent framework.

As the major domain in PISA 2015, work on identifying possible subscales for science, in addition to the overall scientific literacy scale, began with a review of the subscales used in the 2006 cycle, when science was last a major domain. The subscales selected for inclusion in the PISA 2006 database were the three competency-based subscales based on the scientific dimensions documented in the framework: *explaining phenomena scientifically*, *identifying scientific issues* and *using scientific evidence*. The 2015 expert group recommended reporting again on the three scientific competencies, as they were defined in the updated framework: *explain phenomena scientifically*, *evaluate and design scientific enquiry*, and *interpret data and evidence scientifically*. In addition, the expert group recommended that two knowledge subscales be reported: *content knowledge* and *procedural/epistemic knowledge*. Procedural and epistemic knowledge were combined into a single reporting subscale due to a limited number of epistemic items in some of the administered forms. Finally, for continuity with previous reporting scales, three systems – *physical*, *living* and *Earth and space* – were recommended as a third reporting scale.

For reading in the PISA 2000 cycle, in addition to the overall reading literacy scale, two main options were considered: subscales based on the type of reading task and subscales based on the form of reading material. For the international report, the first of these was implemented, leading to the development of subscales to describe the types of reading tasks, or “aspects” of reading: a subscale for *retrieving information*, a second subscale for *interpreting texts* and a third for *reflection and evaluation*. The thematic report for PISA 2000, *Reading for Change*, also reported on the development of subscales based on the form of reading material: *continuous texts* and *non-continuous texts* (OECD, 2002). In the 2009 cycle, volume I of the *PISA 2009 Results* included descriptions of both sets of subscales as well as a combined print reading scale (OECD, 2010). The names of the aspect subscales were modified in order to better apply to digital as well as print reading tasks. The modified aspect category names were *access and retrieve* (replacing *retrieving information*), *integrate and interpret* (replacing *interpreting texts*) and *reflect and evaluate* (for *reflection and evaluation*). For digital reading, a separate, single scale was developed based on the digital reading assessment items administered in 19 countries in PISA 2009 as an international option (OECD, 2011). For PISA 2012, when reading reverted to minor domain status, a single print reading scale was reported, along with a single digital reading scale.

In the case of mathematics, a single mathematical literacy scale was developed for PISA 2000. With the additional data available in the 2003 survey cycle, when mathematics was the major test domain, subscales based on the four overarching ideas – *space and shape*, *change and relationships*, *quantity* and *uncertainty* – were reported. In PISA 2006 and PISA 2009, when mathematics was again a minor domain, only a single scale was reported. For PISA 2012, the expert group carried out a comprehensive revision of the framework at the specific behest of the PISA Governing Board that indicated an interest in seeing mathematical process dimensions used as the primary basis for reporting in mathematics. As well as considering ways in which this could be done, the mathematics expert group also had to consider how the addition of the optional computer-based assessment component included in this cycle could be incorporated into the reporting for 2012. The outcome of these considerations was, firstly, a decision that the computer-based items would be used to expand the same mathematical literacy dimension that was expressed through the paper-based items. Secondly, the expert group recommended that three process-based subscales should be reported. These included: *formulating situations mathematically* (or “*formulate*”), *employing mathematical concepts, facts, procedures and reasoning* (or “*employ*”), and *interpreting, applying and evaluating mathematical outcomes* (or “*interpret*”). In addition, for continuity with the PISA 2003 reporting scales, the content-based scales including *space and shape*, *change and relationships*, *quantity*, and *uncertainty and data* (formerly “*uncertainty*”), were also reported.



For both collaborative problem solving and the optional assessment of financial literacy in PISA 2015, proficiency descriptions on a single overall reporting scale were developed.

## Developing an item map

Based on item performance in the main survey, the test items in the study can be ordered from easiest to most difficult and this range of difficulty can be described using an item map. The item map contains a brief description of a selected number of released items along with their scale values. These descriptions explain the specific skills each item is designed to assess and are linked to the descriptions of performance at each level for the overall scale. As a result, the item map provides some insight into the range of skills and knowledge required of students and the proficiencies they need to demonstrate at various points along the scale.

## DEFINING THE PROFICIENCY LEVELS

The proficiency levels for each of the PISA domains were defined in previous cycles when each was first a major domain. The goal of that process was to decide how to divide the proficiency continuum up into levels that might have some utility. And, having defined those levels, decisions needed to be made about how to decide on the level to which a particular student should be assigned.

The relationship between the observed responses, on the one hand, and student proficiency and item characteristics, on the other hand, is probabilistic. That is, there is some probability that a particular student can correctly solve any particular item and each item can be differentially responsive to the proficiency being measured.

One of the basic tenets of the measurement of human skills or proficiencies is this: If a student's proficiency level exceeds the item's demands, the probability that the student can successfully complete that item is relatively high, and if the student's proficiency is lower than that required by the item, the probability of success for that student on that item is relatively low. The rate of change of the probability of success across the range of proficiency for each item is also affected by the sensitivity of the item to the proficiency scale.

This leads to the question as to the precise criterion that should be used to locate a student on the same scale as that on which the items are located. How can we assign a location that represents student proficiency in meaningful ways? When placing a student at a particular point on the scale, what probability of success should we deem sufficient in relation to items located at the same point on the scale? If a student were given a test comprising a large number of items, each with the same item characteristics, what proportion of those items would we expect the student to successfully complete? Or, thinking of it in another way, if a large number of students of equal ability were given a single test item with a specified item characteristic, about how many of those students would we expect to successfully complete the item?

The answers to these questions depend on assumptions about how items differ in their characteristics or how items function, as well as on what level of probability is deemed a *sufficient probability of success*. In order to define and report PISA outcomes in a consistent manner, an approach is needed to define performance levels and to associate students with those levels. The methodology that was developed and used for previous cycles of PISA was essentially retained for PISA 2015, except that a more general statistical model was used to estimate item parameters, including difficulties (see Chapter 9 for details).

Defining proficiency levels for PISA 2000 progressed in two broad phases. The first, which came after the development of the described scales, was based on a substantive analysis of PISA items in relation to the aspects of literacy that underpinned each test domain. This produced descriptions of increasing proficiency that reflected observations of student performance and a detailed analysis of the cognitive demands of PISA items. The second phase involved decisions about where to set cut-off points for levels and how to associate students with each level in order to lay out how a *sufficient probability of success* plays out in these levels. This is both a technical and a very practical matter of interpreting what it means to be at a level, and has significant consequences for reporting national and international results.

Several principles were considered in developing and establishing a useful meaning of being at a level, and therefore for determining an approach to locating cut-off points between levels and associating students with them. For the levels to provide useful information to PISA stakeholders, it is important to develop a common understanding of what performance at each of those levels means.

First, it is important to understand that the skills measured in each PISA domain fall along a continuum: There are no natural breaking points to mark borderlines between stages along this continuum. Dividing the continuum into levels, though useful for communication about students' development, is essentially arbitrary. Like the definition of units on, for example, a scale of length, there is no fundamental difference between 1 metre and 1.5 metres – it is a matter of degree. It is useful, however, to define stages, or levels along the continua, because they enable us to communicate about the proficiency of students in terms other than continuous numbers. This is a rather common concept, an approach we all know from categorising shoes or shirts by size (S, M, L, XL, etc.).

The approach adopted for PISA 2000 was that it would only be useful to regard students as having attained a particular level if this would mean that we can have certain expectations about what these students are capable of, in general, when they are said to be at that level. It was thus decided that this expectation would have to mean, at a minimum, that students at a particular level would be more likely than not to successfully complete tasks at that level. By implication, it must be expected that they would succeed on at least half of the items on a test composed of items uniformly spread across that level. This definition of being "at a level" is useful in helping to interpret the proficiency of students at different points across the proficiency range defined at each level.

For example, the expectation is that students located at the bottom of a level would complete at least 50% of tasks correctly on a test set at the level, while students at the middle and top of each level would be expected to achieve a higher success rate. At the top border of a level would be the students who have mastered that level. These students would be likely to solve a high proportion of the tasks at that level. But, being at the top border of that level, they would also be at the bottom border of the next highest level where, according to the reasoning here, they should have at least a 50% likelihood of solving any tasks defined to be at that higher level.

Furthermore, the meaning of being at a level for a given scale should be more or less consistent for each level and, indeed, also for scales from the different domains. In other words, to the extent possible within the substantively based definition and description of levels, cut-off points should create levels of more or less constant breadth. Some small variation may be appropriate, but for interpretation and definition of cut-off points and levels to be consistent, the levels have to be about equally broad within each scale. The exception would be the highest and lowest proficiency levels, which are unbounded.

Thus, a consistent approach should be taken to defining levels for the different scales. Their breadth may not be exactly the same for the proficiency scales in different domains, but the same kind of interpretation should be possible for each scale that is developed. This approach links the two variables mentioned in the preceding paragraphs, and third related variable. The three variables can be expressed as follows:

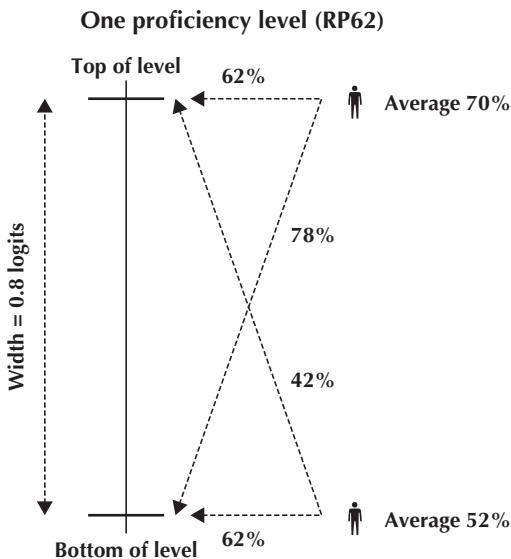
- the expected success of a student at a particular level on a test containing items at that level (proposed to be set at a minimum that is near 50% for the student at the bottom of the level and greater for students who are higher in the level)
- the width of the levels in that scale (determined largely by substantive considerations of the cognitive demands of items at the level and data related to student performance on the items)
- the probability that a student in the middle of a level would correctly answer an item of average difficulty for that level (in fact, the probability that a student at any particular level would get an item at the same level correct), sometimes referred to as the "RP value" for the scale, where "RP" indicates "response probability".

Figure 15.2 summarises the relationship among these three mathematically linked variables under a particular scenario. The vertical line represents a segment of the proficiency scale, with marks delineating the "top of level" and "bottom of level" for any level one might want to consider, with a width of 0.8 logits between the boundaries of the level (noting that this width can vary somewhat for different domains). The RP62 indicates that students will be located on the scale at a point that gives them a 62% chance of getting a typical item at that same level correct.<sup>1</sup> The student represented near the top of the level shown has a 62% chance of getting an item correct that is located at the top of the level, and similarly the student represented at the bottom of the level has the same chance of correctly answering a question at the bottom of the level. A student at the bottom of the level will have an average score of about 52% correct on a set of items spread uniformly across the level. Of course, that student will have a higher likelihood (62%) of getting an item at the bottom of the level correct, and a lower likelihood (about 42%) of getting an item at the top of the level correct. A student at the top of the level will have an average score of about 70% correct on a set of items spread uniformly across the level. That student will have a higher likelihood (about 78%) of getting a typical item at the bottom of the level correct and a lower likelihood (62%) of getting an item at the top of the level correct.



■ Figure 15.2 ■

### Calculating the RP values used to define PISA proficiency levels



PISA 2000 implemented the following solution: Start with the range of described abilities for each bounded level in each scale (the desired band breadth); then determine the highest possible RP value that will be common across domains potentially having bands of slightly differing breadth that would give effect to the broad interpretation of the meaning of being at a level (an expectation of correctly responding to a minimum of 50% of the items in a test comprising items spread uniformly across that level). The value RP = 0.62 is a probability value that satisfied the logistic equations for typical items in that level through which the scaling model is defined, subject to the two constraints mentioned earlier (a width per level of about 0.8 logits and the expectation that a student would get at least half of the items correct on a hypothetical test composed of items spread evenly across the level). In fact, RP=0.62 satisfied the requirements for any scales having band widths up to about 0.97 logits.

The highest and lowest levels are unbounded. For a certain high point on the scale and below a certain low point, the proficiency descriptions could, arguably, cease to be applicable. At the high end of the scale, this is not such a problem since extremely proficient students could reasonably be assumed to be capable of at least the achievements described for the highest level. At the other end of the scale, however, the same argument does not hold. A lower limit therefore needs to be determined for the lowest described level, below which no meaningful description of proficiency is possible. It was proposed that the floor of the lowest described level be set so that it was the same breadth as the other described levels. Student performance below this level is lower than that which PISA can reliably assess and, more importantly, describe.

### REPORTING THE RESULTS FOR PISA SCIENCE

In this section, the ways in which levels of scientific literacy are defined, described and reported will be discussed. This will be illustrated using a subset of items from the PISA 2015 assessment.

### Building an item map for science

The data from the PISA science assessment were analysed to estimate a set of item characteristics for the 184 items included in the main survey.<sup>2</sup> During the process of item development, each item was classified to reflect the scientific competency and type of knowledge it required. In addition, items were classified based on specific content knowledge, or systems (physical systems, living systems or Earth and space systems), as well as their context (personal, local/national or global). Following data analysis, the items were associated with their difficulty estimates and framework classifications. Figure 15.3 shows the item map, which includes this information along with a brief qualitative description for the released items from the PISA 2015 test. Each row in Figure 15.3 represents an individual item. The selected items have

been ordered according to their difficulty, with the most difficult at the top, and the least difficult at the bottom. The difficulty estimate for each item is given, along with the associated classifications and descriptions.

When an item map such as this is prepared, it becomes possible to look for factors that are associated with item difficulty. This can be done by referring to the ways in which scientific literacy is associated with questions located at different points ranging from the bottom to the top of the scale. For example, the item map in Figure 15.3 shows that the easiest items tend to require the application of everyday content knowledge and the ability to recognise aspects of simple scientific phenomena. The most difficult items, by contrast, draw on a range of interrelated scientific ideas and concepts and require the application of sophisticated procedural and epistemic knowledge to offer explanatory hypotheses of novel scientific phenomena, events and processes.

■ Figure 15.3 ■  
A map for selected science items

Code	Item Name	Item Difficulty (RP=0.62)	Item Demands	Explain Phenomena	Evaluate & Design Scientific Enquiry	Interpret Data & Evidence	Content	Procedural	Epistemic	Physical	Living	Earth and Space
CS601Q01	Sustainable Fish Farming	740	Use multiple sources of information to evaluate a system in an unfamiliar context and the interaction among elements in that system.	●			●				●	
CS623Q04	Running in Hot Weather	641	Draw on scientific knowledge to explain a biological reason for an outcome observed in a simulated experiment.	●			●				●	
CS656Q02	Bird Migration	630	Identify a factor that could result in an inadequate or inaccurate set of data and explain its effect.		●			●			●	
CS623Q06	Running in Hot Weather	598	Run a simulated experiment manipulating two independent variables. Use those results to hypothesize the outcome of the experiment with a value for one variable that is not available in the simulation. Select data from the experiment supporting that choice and explain how it does so.		●			●			●	
CS637Q05	Slope-Face Investigation	589	Draw on epistemic knowledge and use provided data to identify the appropriate conclusion from an experiment using controls, providing a reason that justifies that choice.			●			●			●
CS623Q05	Running in Hot Weather	592	Given one defined variable, run a simulated experiment to identify the highest level for a second variable before a negative outcome would occur. Select data from the experiment supporting that choice and explain how it does so.		●			●			●	
CS601Q04	Sustainable Fish Farming	585	Go beyond the provided information to identify a procedure that would meet a specified goal.	●			●			●		
CS623Q02	Running in Hot Weather	580	Run a simulated experiment holding two variables constant and identify the effect of varying the third. Select data from the experiment supporting that choice.			●	●				●	
CS656Q04	Bird Migration	574	Identify one or more statements supported by information provided in two moderately complex representations of data.			●		●			●	
CS623Q03	Running in Hot Weather	531	Given one defined variable, run a simulated experiment to identify the impact of a second variable and identify data supporting that choice.		●			●			●	
CS637Q01	Slope-Face Investigation	517	Draw on epistemic knowledge to explain why a simple experimental design includes two independent measures of a phenomenon.		●				●			●
CS656Q01	Bird Migration	501	Draw on knowledge of life science to identify an explanation of a familiar phenomenon.	●			●				●	
CS623Q01	Running in Hot Weather	497	Follow instructions to carry out and interpret the results of a simple simulated experiment involving two independent variables.			●		●			●	
CS641Q01	Meteoroids & Craters	483	Use simple scientific knowledge to identify the effect of Earth's mass on the speed of objects entering the atmosphere.	●			●			●		
CS601Q02	Sustainable Fish Farming	456	Identify one component of a system that will result in a desired outcome, given an explanation of the function performed by each component.			●	●				●	
CS641Q02	Meteoroids & Craters	450	Use simple scientific knowledge to identify the relationship between a planet's atmosphere and the likelihood that meteoroids will burn up before hitting the planet surface.	●			●				●	
CS641Q04	Meteoroids & Craters	438	Use familiar and simple scientific knowledge to order three craters by their age from oldest to newest, based on an image showing craters of different sizes.			●	●				●	
CS641Q03	Meteoroids & Craters	299	Use everyday scientific knowledge to match the size of a meteoroid with the size of the crater it would create on a planet's surface, based on an image showing three craters of different sizes.			●	●				●	

Based on the patterns observed in the science item pool, it was possible to characterise the increasing complexity of competencies measured. This can be done by referring to the ways in which science competencies are associated with items located at different points, ranging from the bottom to the top of the scale. The ascending difficulty of science questions in PISA 2015 is associated with the following attributes, which require all three competencies but shift in



emphasis as students progress from the application of simple everyday knowledge to using more sophisticated content, procedural and epistemic knowledge to develop hypotheses about novel scientific phenomena, events and processes.

The attributes include the following:

- The degree to which the transfer and application of knowledge is required. At the lowest levels the application of knowledge is simple and direct. The requirement can often be fulfilled with simple recall of single facts. At higher levels of the scale, individuals are required to draw on multiple fundamental concepts and combine categories of knowledge in order to respond correctly.
- The degree of cognitive demand required to analyse the presented situation and to synthesise an appropriate answer. The 2015 Scientific Literacy framework defined increasing complexity based on levels of cognitive demand within the assessment of scientific literacy and across all three competencies of the framework. The factors that determine the cognitive demand of items in science include: the number of elements of knowledge and their degree of complexity; the level of familiarity and prior knowledge that students may have of the content, procedural and epistemic knowledge involved; the cognitive operation required by the item (e.g., recall, analysis, evaluation); and the extent to which forming a response is dependent on models or abstract scientific ideas.

For example, items with low cognitive complexity typically involve carrying out a one-step procedure, for example, recalling a fact, term, principle, or concept, or locating a single point of information from a graph or table. Items with medium cognitive complexity require the use and application of conceptual knowledge to describe or explain phenomena, select appropriate procedures involving two or more steps, organise or display data, or interpret or use simple data sets or graphs. Finally, items with high cognitive demand require students to analyse complex information or data, synthesise or evaluate evidence, reason given various sources, or develop a plan or sequence of steps to approach a problem.

- The degree of analysis needed to answer a question is also an important driver of difficulty. This includes the demands arising from the requirement to discriminate among issues presented in the situation under analysis, identify the appropriate knowledge domain, and use appropriate evidence for claims or conclusions. The analysis may include the extent to which the scientific demands of the situation are clearly apparent or whether students must differentiate among components of the situation to clarify the scientific issues as opposed to other non-salient or non-scientific issues.
- The degree of synthesis required may impact item complexity. Synthesis may range from a single piece of evidence where no real construction of justification or argument is required to situations requiring students to apply multiple sources of evidence and compare competing lines of evidence and different explanations to adequately argue a position.

### Defining levels of scientific literacy

The reporting approach used by the OECD has been defined in previous cycles of PISA and is based on the definition of a number of levels of proficiency. Descriptions were developed to characterise typical student performance at each level. The levels were used to summarise the performance of students, to compare performances across subgroups of students, and to compare average performances among groups of students, in particular among the students from different participating countries. A similar approach has been used here to analyse and report PISA 2015 outcomes for science.

For PISA 2006 science, student scores were transformed to the PISA scale, with a mean of 500 and a standard deviation of 100, and levels of proficiency were defined and described. In accordance with the approach taken for the other PISA domains, the science scale has been extended to describe one level below the lowest previously-described level. Thus the PISA 2015 science scale has seven described levels instead of the six defined for PISA 2006. The previously-named Level 1 was renamed Level 1a and the level defined below this was named Level 1b.

The level definitions on the PISA scale are given in Table 15.1.

**Table 15.1 Scientific literacy performance band definitions on the PISA scale**

Level	Score points on the PISA scale
6	Higher than 707.93
5	Higher than 633.33 and less than or equal to 707.93
4	Higher than 558.73 and less than or equal to 633.33
3	Higher than 484.14 and less than or equal to 558.73
2	Higher than 409.54 and less than or equal to 484.14
1a	Higher than 334.94 and less than or equal to 409.54
1b	260.54 to less than or equal to 334.94

Information about the items in each level is used to develop summary descriptions of the kinds of scientific literacy associated with different levels of proficiency. These summary descriptions can then be used to encapsulate typical science proficiency of students associated with each level. As a set, they describe development in scientific literacy.

PISA is administered once every three years, with each of the three core domains the major focus in turn. Science was the major domain in PISA 2006. PISA 2015, therefore, had a set of level descriptors upon which to build. The new items that were developed for PISA 2015 were considered in relation to the existing level descriptions and in relation to the preliminary descriptions that were included in the 2015 framework for scientific literacy. The focus was first on the descriptions for the overall science scale, presented here in Figure 15.4.

■ Figure 15.4 ■

#### **Summary descriptions of the seven proficiency levels on the scientific literacy scale**

Level	What students can typically do
6	At Level 6, students can draw on a range of interrelated scientific ideas and concepts from the physical, life and Earth and space sciences and use procedural and epistemic knowledge in order to offer explanatory hypotheses of novel scientific phenomena, events and processes that require multiple steps or to make predictions. In interpreting data and evidence, they are able to discriminate between relevant and irrelevant information and can draw on knowledge external to the normal school curriculum. They can distinguish between arguments that are based on scientific evidence and theory and those based on other considerations. Level 6 students can evaluate competing designs of complex experiments, field studies or simulations and justify their choices.
5	At Level 5, students can use abstract scientific ideas or concepts to explain unfamiliar and more complex phenomena, events and processes. They are able to apply more sophisticated epistemic knowledge to evaluate alternative experimental designs and justify their choices and use theoretical knowledge to interpret information or make predictions. Level 5 students can evaluate ways of exploring a given question scientifically and identify limitations in interpretations of data sets including sources and the effects of uncertainty in scientific data.
4	At Level 4, students can use more sophisticated content knowledge, which is either provided or recalled, to construct explanations of more complex or less familiar events and processes. They can conduct experiments involving two or more independent variables in a constrained context. They are able to justify an experimental design, drawing on elements of procedural and epistemic knowledge. Level 4 students can interpret data drawn from a moderately complex data set or less familiar contexts and draw appropriate conclusions that go beyond the data and provide justifications for their choices.
3	At Level 3, students can draw upon moderately complex content knowledge to identify or construct explanations of familiar phenomena. In less familiar or more complex situations, they can construct explanations with relevant cueing or support. They can draw on elements of procedural or epistemic knowledge to carry out a simple experiment in a constrained context. Level 3 students are able to distinguish between scientific and non-scientific issues and identify the evidence supporting a scientific claim.
2	At Level 2, students are able to draw on everyday content knowledge and basic procedural knowledge to identify an appropriate scientific explanation, interpret data, and identify the question being addressed in a simple experimental design. They can use everyday scientific knowledge to identify a valid conclusion from a simple data set. Level 2 students demonstrate basic epistemic knowledge by being able to identify questions that could be investigated scientifically.
1a	At Level 1a, students are able to use everyday content and procedural knowledge to recognise or identify explanations of simple scientific phenomenon. With support, they can undertake structured scientific enquiries with no more than two variables. They are able to identify simple causal or correlational relationships and interpret graphical and visual data that require a low level of cognitive demand. Level 1a students can select the best scientific explanation for given data in familiar personal, local and global contexts.
1b	At Level 1b, students can use everyday content knowledge to recognise aspects of simple scientific phenomenon. They are able to identify simple patterns in data, recognise basic scientific terms and follow explicit instructions to carry out a scientific procedure.

Figures 15.5, 15.6 and 15.7 provide the summary descriptions of knowledge and skills required to complete tasks located within the defined bands for the three competency subscales: *explaining phenomena scientifically*, *evaluating and designing scientific enquiry* and *interpreting data and evidence scientifically* respectively.

■ Figure 15.5 [Part 1/2] ■

#### **Summary descriptions of the proficiency levels on the scientific literacy subscale Explain phenomena scientifically**

Level	General proficiencies students should have at each level	Tasks a student should be able to do
6	At Level 6, students can draw on a range of inter-related scientific ideas and concepts from life, physical or Earth and space sciences to make predictions or to construct explanations of novel and unfamiliar phenomena, events and processes that may involve several steps. They can demonstrate the use of knowledge beyond standard science curricula and use procedural and epistemic knowledge appropriately.	<ul style="list-style-type: none"> <li>■ Construct acceptable scientific explanations, using a broad range of knowledge, ideas and concepts.</li> <li>■ Recognise when data/information in the text does not answer the question.</li> <li>■ Use given scientific knowledge and recall additional relevant scientific knowledge to explain an unfamiliar phenomenon.</li> <li>■ Construct and run a mental model to offer an explanation or make a prediction in an unfamiliar situation.</li> <li>■ Comment on the appropriate use of scientific models and their limitations.</li> </ul>
5	Students at this level can use abstract scientific ideas or concepts to explain more complex phenomena, events and processes, which may be unfamiliar.	<ul style="list-style-type: none"> <li>■ Select an appropriate scientific explanation of an unfamiliar event, phenomenon or process.</li> <li>■ Construct an appropriate explanation drawing upon abstract scientific ideas and constructs.</li> <li>■ Apply theoretical scientific knowledge to interpret given information, develop an explanation or make a prediction.</li> </ul>



■ Figure 15.5 [Part 2/2] ■

**Summary descriptions of the proficiency levels on the scientific literacy subscale  
Explain phenomena scientifically**

Level	General proficiencies students should have at each level	Tasks a student should be able to do
4	At Level 4, students can recall or use given scientific ideas to construct explanations of relatively complex or less familiar events and processes, or to make simple predictions.	<ul style="list-style-type: none"> <li>■ Identify or construct an appropriate causal explanation for a more complex or less familiar phenomenon, event or process.</li> <li>■ Identify the relationship between simple physical quantities and use this to explain a phenomenon.</li> <li>■ Predict how one quantity will change when other quantities change.</li> <li>■ Use scientific knowledge to evaluate a claim or to interpret an unfamiliar phenomenon.</li> <li>■ Recognise relationships between physical quantities.</li> </ul>
3	Students at this level can draw upon moderately complex scientific facts and ideas to identify or construct appropriate simple explanations of familiar phenomena. In less familiar or more complex situations, they can construct an explanation with relevant cueing or support.	<ul style="list-style-type: none"> <li>■ Construct simple explanations of familiar phenomena drawing on knowledge from life, physical or Earth and space sciences.</li> <li>■ Identify a conclusion consistent with given information in an unfamiliar context.</li> <li>■ Select from multiple components and place them in a logical order to construct simple explanations.</li> <li>■ Identify causal factors which explain a phenomenon.</li> </ul>
2	At Level 2, students can recall and apply simple scientific facts and ideas, or select a simple scientific explanation, given relevant cues and support.	<ul style="list-style-type: none"> <li>■ Use familiar and simple scientific knowledge to draw an appropriate conclusion.</li> <li>■ Select the correct explanation of a relatively familiar scientific situation.</li> <li>■ Choose appropriate alternatives to complete an explanation.</li> <li>■ Use simple scientific knowledge to identify causal relationships.</li> <li>■ Reconstruct a temporal sequence for a familiar scientific phenomenon.</li> </ul>
1a	Students at this level can select an appropriate example of a given simple scientific concept or identify an appropriate scientific explanation for a familiar event or process that is consistent with given information.	<ul style="list-style-type: none"> <li>■ Use familiar content and procedural knowledge to recognise or identify explanations.</li> <li>■ Select the best scientific explanation from a list for given data in familiar contexts.</li> </ul>
1b	Students at this level can recognise scientific terms and use single scientific facts close to their personal experience to recognise very simple cause and effect relationships.	<ul style="list-style-type: none"> <li>■ Recognise simple scientific language or scientific conventions used in everyday life situations.</li> <li>■ Use familiar content knowledge to recognise scientific aspects of simple phenomena in tasks that require a low level of cognitive demand.</li> </ul>

■ Figure 15.6 [Part 1/2] ■

**Summary descriptions of the seven proficiency levels on the scientific literacy subscale  
Evaluate and design scientific enquiry**

Level	General proficiencies students should have at each level	Tasks a student should be able to do
6	At Level 6, students can evaluate competing designs of complex experiments, field studies or simulations and justify their choices.	<ul style="list-style-type: none"> <li>■ Evaluate an investigation involving multiple variables requiring the identification of the independent or dependent variable.</li> <li>■ Justify choices and the range of data to be collected drawing on relevant epistemic and/or procedural knowledge.</li> <li>■ Evaluate and comment on the model inherent to experimental designs.</li> </ul>
5	Students at this level can evaluate alternative experimental designs or data interpretations and justify their choices. They can identify limitations of the interpretations of data sets.	<ul style="list-style-type: none"> <li>■ Evaluate whether an empirical question can be answered scientifically or not.</li> <li>■ Justify a more detailed feature of an experimental design.</li> <li>■ Provide a procedural justification for the inadequacy of a set of data.</li> <li>■ Choose between two experimental designs and justify the choice drawing on procedural, epistemic or content knowledge.</li> <li>■ Justify a data collection procedure in a context involving several independent variables.</li> </ul>
4	At Level 4, students can conduct experiments involving two or more independent variables in a constrained context and justify aspects of their experimental design, drawing on procedural and epistemic knowledge. They can interpret data drawn from more complex or less familiar contexts and draw appropriate conclusions that go beyond the data.  They can use data from less familiar contexts to identify trends and make predictions.	<ul style="list-style-type: none"> <li>■ Carry out and interpret a simple experiment involving the manipulation of more than one independent variable.</li> <li>■ Follow instructions to identify the outcome of several variable choices.</li> <li>■ Manipulate variables to answer a scientific question, identify a trend, interpolate between, or extrapolate beyond, the data.</li> <li>■ Justify the conclusions of an experimental design drawing on procedural or epistemic knowledge.</li> <li>■ Identify the question of an investigation of a more complex or less familiar experimental design.</li> </ul>

■ Figure 15.6 [Part 2/2] ■

**Summary descriptions of the seven proficiency levels on the scientific literacy subscale  
Evaluate and design scientific enquiry**

Level	General proficiencies students should have at each level	Tasks a student should be able to do
3	Students at this level can draw on procedural or epistemic knowledge to design, and justify aspects of the design of, a simple experiment in a constrained context. They can distinguish between scientific, technological and non-scientific issues.	<ul style="list-style-type: none"> <li>■ Identify which variable to control in a two variable experiment.</li> <li>■ Drawing on epistemic or procedural knowledge, provide a justification for aspects of a simple experimental design.</li> <li>■ Identify the role of simulations in scientific enquiry.</li> <li>■ Discriminate between issue that can be solved by science or other means.</li> <li>■ Within a constrained context, identify a set of data that could answer a specified question about a phenomenon.</li> </ul>
2	Students at Level 2 are able to draw on procedural and basic content knowledge to identify the question being addressed in a simple experimental design. They can collect and interpret data to answer questions that require only simple or everyday content knowledge. They can distinguish between a non-scientific and scientific question.	<ul style="list-style-type: none"> <li>■ Given a simple experimental design, identify the question being addressed.</li> <li>■ Distinguish between simple scientific and simple non-scientific questions.</li> <li>■ Interpret simple data sets and draw an appropriate conclusion using everyday knowledge.</li> <li>■ Carry out a straightforward procedure to collect a data set to answer a simple question.</li> <li>■ Identify the aspects of a simple model and the external features they represent.</li> </ul>
1a	Students at this level can, with support, carry out a simple experiment involving one independent and one dependent variable to generate data to answer a question.	<ul style="list-style-type: none"> <li>■ Identify the independent variable in a given situation.</li> <li>■ Follow instructions to carry out a simple experiment to investigate how an outcome changes when one independent variable is changed.</li> </ul>
1b	At this level, students typically are able to follow simple instructions to carry out a scientific procedure.	<ul style="list-style-type: none"> <li>■ Run a simulation to extract a single data point.</li> </ul>

■ Figure 15.7 ■

**Summary descriptions of the seven proficiency levels on the scientific literacy subscale  
Interpret data and evidence scientifically**

Level	General proficiencies students should have at each level	Tasks a student should be able to do
6	At Level 6, students can evaluate the strength of support provided by data for competing hypotheses and construct and justify a conclusion using abstract science concepts. They can also discriminate between relevant and irrelevant information, and draw on outside knowledge to construct an explanation.	<ul style="list-style-type: none"> <li>■ Evaluate a complex set of data to determine whether each piece of data supports one, both or neither of two or more competing hypotheses.</li> <li>■ Provide a reason for their choice using abstract science concepts and applying procedural or epistemic knowledge.</li> </ul>
5	Students at this level can interpret a moderately complex data set to construct and justify a conclusion using abstract science concepts. They can also identify sources and effects of uncertainty in scientific data.	<ul style="list-style-type: none"> <li>■ Analyse complex data to identify which of several inferences is correct.</li> <li>■ Generate a set of data from a simulation by manipulating a single variable to identify the correct outcome from a number of possibilities.</li> </ul>
4	At Level 4, students can interpret and manipulate a moderately complex data set expressed in a number of formats to select or justify appropriate conclusions. They can also distinguish between scientific and social or personal issues when interpreting data.	<ul style="list-style-type: none"> <li>■ Analyse moderately complex data to identify which of several inferences is correct.</li> <li>■ Analyse more complex data to identify the appropriate conclusion of an experiment using controls and provide a reason that justifies their choice.</li> </ul>
3	Students at this level can interpret and transform data to support a claim or conclusion. They can identify the evidence supporting a scientific claim.	<ul style="list-style-type: none"> <li>■ Analyse a data table to identify which of several inferences is correct.</li> <li>■ Use data to identify the appropriate conclusion from an experiment using controls or a set of data and provide a reason that justifies their choice.</li> </ul>
2	Students at this level can identify data that support a claim or conclusion and interpret data to select relevant explanations.	<ul style="list-style-type: none"> <li>■ Analyse tabular or graphic data to identify which of several hypotheses or claims are supported by the data.</li> <li>■ Identify the pattern in a data set such as a graph or table.</li> </ul>
1a	Students at this level can identify whether simple data support a claim or conclusion. They can make straightforward interpretations of simple data sets presented in different formats.	<ul style="list-style-type: none"> <li>■ Identify the trend in simple data set.</li> <li>■ Transform simple data representations between pictorial, graphical, tabular and text.</li> <li>■ Use a simple data set to Identify data that support a conclusion.</li> </ul>
1b	Students at this level can identify simple patterns in data.	<ul style="list-style-type: none"> <li>■ In response to a specific question showing a simple pictorial representation of objects, make comparisons and judgments about the differences observed.</li> </ul>



## **Notes**

1. A typical item is one that has an item slope parameter of 1.0 in this example. In the more general statistical model used for 2015, items are allowed to vary in their slope parameter, which quantifies the strength of the relationship between proficiency and item response. This slope parameter was introduced to allow tasks to be more appropriately described if not fitted well by assuming a common slope for all items. This leads to a reporting model that described the observed student responses much more appropriately but, as a consequence, it requires talking about typical items in terms of RP62 and proficiency levels.
2. For a detailed description of the scaling procedures used in PISA 2015, see Chapter 9 of this report.

## **References**

- OECD** (2017), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264281820-en>.
- OECD** (2011), *PISA 2009 Results: Students On Line: Digital Technologies and Performance (Volume VI)*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264112995-en>.
- OECD** (2010), *PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264091450-en>.
- OECD** (2002), *Reading for Change: Performance and Engagement across Countries: Results from PISA 2000*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264099289-en>.





16

# Scaling procedures and construct validation of context questionnaire data

<b>Introduction .....</b>	290
<b>Scaling methodology and construct validation .....</b>	290
<b>School Questionnaire derived variables .....</b>	321
<b>Educational Career Questionnaire .....</b>	327
<b>ICT Familiarity Questionnaire .....</b>	328
<b>Parent Questionnaire .....</b>	331
<b>Teacher Questionnaires .....</b>	335
<b>The PISA index of economic, social and cultural status (ESCS) .....</b>	339

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.



## INTRODUCTION

The PISA 2015 Context Questionnaires are based on the questionnaire framework (OECD, 2017), described in Chapter 3 of this report. The questionnaires include numerous indicators for reporting over time (trend indicators) or were designed to be used in analyses as single items (for example, gender). However, many questionnaire items were designed to be combined in some way in order to measure latent constructs that cannot be observed directly (e.g., a student's achievement motivation or economic, social and cultural background). To these items, transformations or scaling procedures were applied to construct meaningful indices.

In the following, these indices are referred to as 'derived variables'. Many derived variables were taken from previous PISA cycles without change as part of the trend content. This chapter describes derived variables based on one or more items that were constructed and validated for all questionnaires administered in PISA 2015.

In analogy to previous PISA surveys, three different kinds of derived variables can be distinguished:

- simple questionnaire indices constructed through the arithmetical transformation or recoding of one or more items
- derived variables based on IRT scaling (see section "Scaling procedures" in this chapter)
- ESCS composite scores (see section "The PISA index of economic, social and cultural status (ESCS)" in this chapter).

As described in Chapter 3, the PISA 2015 Context Questionnaires included a broad scope of context factors assessed with different questionnaire instruments. While student and school context questionnaires were mandatory in all countries, many countries also administered the optional questionnaire to parents of the tested students. In addition, countries could choose to administer the international options *Information and Communication Technology (ICT) Familiarity Questionnaire* and the *Educational Career Questionnaire* to students. Moreover, several countries chose to participate in the *Teacher Questionnaire* option including questionnaires for science and non-science teachers (See Chapter 17 for an overview of participation in international options).

This chapter (i) describes the methodology used for scaling and construct validation including trend scales, (ii) presents an overview of all derived variables (simple indices, IRT-based scales) per questionnaire, and (iii) illustrates the computation of the *PISA index of economic, social and cultural status (ESCS)*.

## SCALING METHODOLOGY AND CONSTRUCT VALIDATION

### Scaling procedures

As in previous cycles of PISA, one subset of the derived variables was constructed using IRT (*item response theory*) scaling methodology. In the IRT framework, a number of different models can be distinguished with the generalised partial credit model (see below) being the one used for constructing derived variables in the PISA 2015 Context Questionnaires.

For each item, item responses are modelled as a function of the latent construct,  $\theta_j$ . With the one-parameter model (*Rasch model*; Rasch, 1960) for dichotomous items, the probability of person  $j$  selecting category 1 instead of 0 is modelled as:

**16.1**

$$P(X_{ji} = 1 | \theta_j, \beta_i) = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)}$$

where  $P(X_{ji} = 1)$  is the probability of person  $j$  to score 1 on item  $i$ ,  $\theta_j$  is the estimated latent trait of person  $j$  and  $\beta_i$  the estimated location or difficulty of item  $i$  on this dimension<sup>1</sup>. In the case of items with more than two ( $m$ ) categories (e.g., Likert-type items), this model can be generalised to the *Partial Credit Model* (Masters and Wright, 1997), which takes the form of:

**16.2**

$$P(X_{ji} = k | \theta_j, \beta_i, d_i) = \frac{\exp\left(\sum_{r=0}^k \theta_j - (\beta_i + d_{ir})\right)}{\sum_{u=0}^m \exp\left(\sum_{r=0}^u \theta_j - (\beta_i + d_{ir})\right)}$$



where  $P(X_{ji} = k)$  denotes the probability of person  $j$  to score  $k$  on item  $i$  out of the  $m_i$  possible scores on the item.  $\theta_j$  denotes the person's latent trait, the item parameter  $\beta_i$  gives the general location or difficulty of the item on the latent continuum and  $d_{ir}$  denote additional step parameters. This model has been used throughout previous cycles of PISA for scaling derived variables of the context questionnaires. However, research literature (especially, Glass and Jehangir, 2014) suggests that a generalisation of this model, the *generalised partial credit model* (GPCM) (Muraki, 1992), is more appropriate in the context of PISA since it allows for the item discrimination to vary between items within any given scale. This model takes the form of:

### 16.3

$$P(X_{ji} = k | \theta_j, \beta_i, \alpha_i, d_i) = \frac{\exp\left(\sum_{r=0}^k \alpha_i (\theta_j - (\beta_i + d_{ir}))\right)}{\sum_{u=0}^{m_i} \exp\left(\sum_{r=0}^u \alpha_i (\theta_j - (\beta_i + d_{ir}))\right)}$$

in which the additional discrimination parameter  $\alpha_i$  allows for the items of a scale to contribute with different weights to the measurement of the latent construct.

Most of the scales were analysed based on 2015 data only (see section “Regular scales”) and other, mostly science-related scales were analysed to allow for comparisons with the weighted likelihood estimates (WLE; Warm, 1989) obtained in PISA 2006 (see section “Trend scales”).

### Regular scales (PISA 2015)

For the regular scales, international item and person parameters were obtained from a GPCM (see formula 16.3) in a single analysis based on data from all persons in all countries using the *mdltm* software (von Davier, 2008). For each scale, only persons with a minimum number of three valid responses were included. Students were weighted using the final student weight (W\_FSTUWT), and all countries contributed equally to the estimation. Additional analyses on the invariance of item parameters across countries and languages were conducted and unique parameters were assigned if necessary (see section “Cross-country comparability” in this chapter). Once this process was finished, weighted likelihood estimates (WLE; Warm, 1989) were used as individual participant scores and transformed to an international metric with an OECD mean of zero and an OECD standard deviation of one<sup>2</sup>. The transformation was achieved by applying formula 16.4:

### 16.4

$$\theta'_j = \frac{\theta_j - \bar{\theta}_{OECD}}{\sigma_{\theta(OECD)}}$$

where  $\theta'_j$  are the WLE scores in the final metric,  $\theta_j$  the original WLEs in logits,  $\bar{\theta}_{OECD}$  is the OECD mean of logit scores with equally weighted country samples, and  $\sigma_{\theta(OECD)}$  is the corresponding OECD standard deviation of the original WLEs. OECD means and standard deviations (S.D.) used for the transformation into the final metric are displayed in Table 16.1.

[Part 1/2]

**OECD mean and standard deviation (S.D.) for the untransformed WLEs of regular scales**

Table 16.1 **Student-level indices** in the different PISA 2015 context questionnaires

Derived variable	N	Mean	S.D.
<b>Student-level indices</b>			
ADINST	149 283	-0.12	1.18
ANXTEST	174 845	0.05	1.03
AUTICT	137 606	0.56	1.31
BELONG	169 366	0.74	1.09
COMPCT	137 619	0.70	1.42
COOPERATE	174 239	0.78	0.99
CPSVALUE	174 095	0.63	1.13
CULTPOSS	174 162	0.05	0.99
DISCLISCI	156 129	0.53	1.30
EMOSUPPS	170 303	1.61	1.26
ENTUSE	142 249	-0.01	0.49

[Part 2/2]

**OECD mean and standard deviation (S.D.) for the untransformed WLEs of regular scales  
in the different PISA 2015 context questionnaires**

Table 16.1

Derived variable	N	Mean	S.D.
EPIST	161 707	0.78	1.22
HEDRES	176 212	1.13	0.78
HOMEPOS	177 199	0.66	0.53
HOMESCH	139 325	-0.56	0.83
IBTEACH	154 036	-0.26	0.82
ICTRES	176 248	0.47	0.79
INTBRSCI	162 260	-0.09	1.11
INTICT	138 858	0.63	1.02
MOTIVAT	174 489	0.81	1.09
PERFEED	151 719	-0.78	1.53
SOIAICT	136 493	-0.01	1.33
TDTEACH	152 358	0.12	1.07
TEACHSUP	154 354	0.75	1.24
USESCH	139 842	-0.83	0.83
WEALTH	176 453	0.70	0.64
<b>School-level indices</b>			
EDUSHORT	168 744	-0.61	1.44
LEAD	167 885	0.32	0.58
LEADCOM	167 632	0.08	0.78
LEADINST	164 939	0.26	0.74
LEADPP	164 777	0.77	0.87
LEADTCH	164 740	0.54	0.90
STAFFSHORT	168 178	-0.72	0.81
STUBEHA	167 746	-0.69	0.97
TEACHBEHA	167 674	-0.96	0.97
<b>Teacher-level indices</b>			
COLSCIT	127 795	0.89	1.27
EXCHT	246 628	0.40	0.64
SATJOB	374 474	1.56	1.45
SATTEACH	375 540	0.99	1.22
SECONT	115 562	1.61	1.13
SETEACH	115 569	1.41	1.14
TCEDUSHORT	372 293	-0.63	1.48
TCLEAD	246 604	0.87	1.72
TCSTAFFSHORT	370 074	-0.74	0.89
<b>Parent-level indices</b>			
CURSUPP	472 202	0.06	0.54
EMOSUPP	469 931	2.31	1.14
PASCHIPOL	465 559	0.53	1.15
PRESUPP	470 030	-0.69	0.68

Note: N reflects the sample size after senate weights were applied. Senate weights were constructed to sum up to the target sample size of 5 000 within each country.

### Trend scales (PISA 2006 - PISA 2015)

For those scales administered in both PISA 2006 and PISA 2015, scale scores in PISA 2015 were constructed to allow for comparisons with those reported in PISA 2006 using a *common calibration linking procedure*. This procedure consists of two phases: calibration and linking phase.

In the calibration phase, international item and person parameters were obtained from a generalised partial credit model (see formula 16.3) in a single analysis based on data from all persons in all countries from both cycles (2006 and 2015) using the *mdltm* software (von Davier, 2008). For each scale, only persons with a minimum number of three valid responses were included. Students were weighted using the final student weight, and each country in each cycle contributed equally to the estimation. Additional analyses on the invariance of item parameters across countries, languages and cycles were conducted and unique parameters were assigned if necessary (see section “Cross-country comparability” in this chapter). WLEs resulting from this concurrent calibration were derived for examinees from both cycles ( $WLE_{2006,new}$ ,  $WLE_{2015}$ ).

In the linking phase, the 2015 WLEs obtained in the calibration phase ( $WLE_{2015}$ ) were linked to the 2006 metric to obtain final WLEs ( $WLE^*_{2015}$ ) by a linear transformation of the following form:

**16.5**

$$WLE^*_{2015} = A \times WLE_{2015} + B$$



The linking constants ( $A$ ,  $B$ ) were calculated based on the mean and standard deviation of the newly derived and original WLEs of the 2006 data:

**16.6**

$$A = \frac{SD_{WLE\ 2006.\ original}}{SD_{WLE\ 2006.\ new}}$$

**16.7**

$$B = M_{WLE\ 2006.\ original} - A \times M_{WLE\ 2006.\ new}$$

Table 16.2 shows both the transformation constants ( $A$ ,  $B$ ) and the correlations between the original and newly derived WLEs for PISA 2006 ( $r(WLE_{2006.\ original}, WLE_{2006.\ new})$ ). They indicate that original and transformed scales are highly consistent both with respect to distributional characteristics and rank order of individuals, indicating that all scales could be recovered well. This is particularly noteworthy as the scaling model changed from the partial credit model in previous cycles of PISA to the generalised partial credit model in 2015.

**Scaling constants ( $A$ ,  $B$ ) and correlations between original and newly derived 2006 WLEs**  
Table 16.2 for trend scales in 2015

Derived variable	B	A	$r(WLE_{2006.\ original}, WLE_{2006.\ new})$
<b>Student-level indices</b>			
ENVWARE	1.05	-0.52	0.991
ENVOPT	1.22	0.66	0.998
INSTCIE	0.56	-0.19	0.999
JOYSCIE	0.58	-0.12	0.998
SCIEACT	0.85	1.29	0.997
SCIEEFF	1.34	-0.33	1.000
<b>Parent-level indices</b>			
PQENPERC	1.64	-2.34	0.999
PQENVOPT	1.11	1.00	0.999
PQGENSCI	0.77	-1.11	0.996
PQSCHOOL	0.82	-0.69	0.995

## Interpreting results from IRT scaling

### Interpreting person parameters

As in previous cycles of PISA, in PISA 2015 categorical items from the context questionnaires were scaled using IRT modelling. WLEs for the latent dimensions were transformed to scales with a mean of 0 and a standard deviation of 1 across OECD countries (with equally weighted countries), meaning that the average OECD student would have an index value of zero and about two-thirds of the OECD student population would be between the values of -1 and 1.

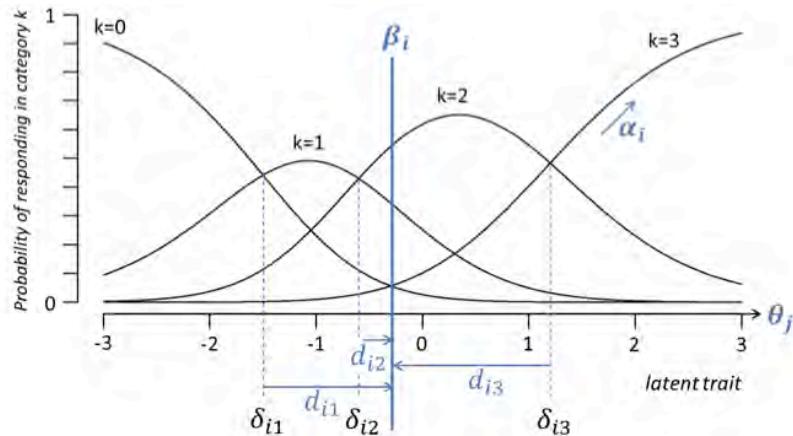
It is possible to interpret these scores by comparing individual scores or group mean scores to the OECD mean, but the individual scores do not reveal anything about the actual item responses and it is impossible to determine from scale score values to what extent respondents endorsed the items used for the measurement of the latent construct. Negative values on the index do not imply that students responded negatively to the underlying question. Rather, students with negative scores are those who responded less positively than the average student across OECD countries. Likewise, students with positive scores are those who responded more positively than the average student in OECD countries.

### Interpreting item parameters

The generalised partial credit model (see formula 16.3) contains three kinds of item parameters: one relating to the general location or difficulty of the item ( $\beta$ ), one relating to the deviance of each of the single response categories from this location parameter ( $d$ ), and one relating to the item's discrimination or slope ( $a$ ). Figure 16.1 displays the category characteristic curves of a four-category item (e.g., a Likert-type item with response categories "strongly disagree", "disagree", "agree", and "strongly agree"). The three kinds of generalised partial credit model item parameters were included in this representation, and each will be discussed in detail below.

■ Figure 16.1 ■

**Item characteristic curves for a four-category item under the generalised partial credit model (GPCM)**  
*Model parameters are highlighted in blue*



The overall item location or difficulty parameter,  $\beta$ , can be regarded as the item's location on the latent continuum of the construct to be measured. The  $m-1$  threshold parameters,  $d$ , of an  $m$ -category item represent deviations from this general location. Thus, the threshold parameters' means equal 0. This parameterization has also been referred to as the *expanded parametrisation* (Penfield, Myers and Wolfe, 2008) and was reported throughout previous cycles of PISA. Combining the location parameter and the  $m$  threshold parameters leads to a reduced parameterization that might be more familiar to some users (e.g. Muraki, 1992). Threshold parameters,  $d$ , and step parameters,  $\delta$ , can easily be converted into each other by:

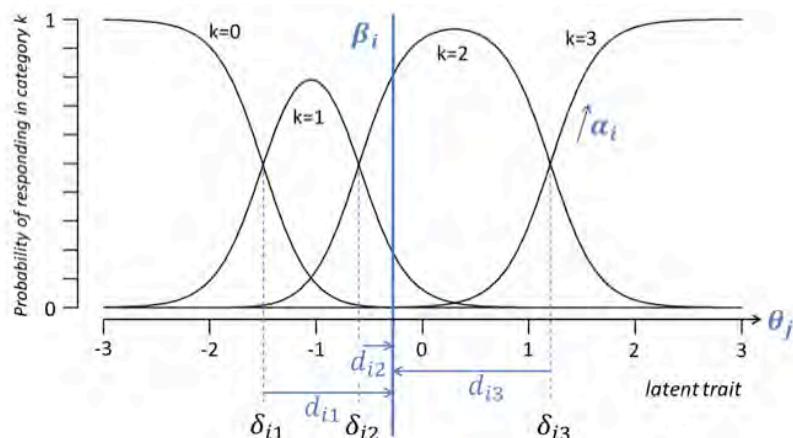
**16.8**

$$\delta_k = \beta - d_k$$

These step parameters,  $\delta$ , signify the intersections between two neighbouring category characteristic curves and thus, the point on the latent continuum at which a response in the higher category becomes more likely. The slope parameter,  $\alpha$ , signifies the slope of the category characteristic curves, thus indicating how well a response in a certain category discriminates between persons on the latent continuum. Figure 16.2 contains category characteristic curves for which only the slope has been increased while holding all other model parameters identical with the model displayed in Figure 16.1. The same increment on the latent continuum leads to a better prediction of the given response.

■ Figure 16.2 ■

**Illustration of an increase of the slope parameter,  $\alpha$ , on category response curves for a four-category item under the generalised partial credit model (GPCM)**





The respective estimates for all three kinds of parameters will be reported along with each item's wording in the subsequent sections. The model parameters can be used to compare the items of a scale with each other: items with a higher overall difficulty are less often "solved", meaning that persons tended to respond in lower categories, and the step parameters shed light on the relative difficulty of the response categories. Items with a higher slope can be seen as better indicators of the latent construct, and, thus, are more represented in the meaning of the scale score (WLE).

In general, the item difficulty parameters of an IRT model can be interpreted with respect to the person parameter,  $\theta$ , and vice versa. Please note that this is not possible in this context, because instead of the original  $\theta$  estimates (WLEs) either standardised values are reported (in case of regular scales) or scores are linked to another scale (in case of trend scales) so that the WLEs are no longer on their original metric.

## Construct validation

The development of comparable measures of student background, practices, attitudes and perceptions is a major goal of PISA. Cross-country validity of these constructs is of particular importance as measures derived from questionnaires are often used to predict differences in student performance within and across countries and are, thus, potential sources of policy-relevant information about ways of improving educational systems. Different methodological approaches for validating questionnaire constructs have been developed. The two approaches implemented for context questionnaires in PISA 2015 are introduced below.

### Internal consistency

Cronbach's alpha was used to check the internal consistency of each scale within the countries and to compare it between the countries. The coefficient ranges between 0 and 1, with higher values indicating higher internal consistency. Commonly accepted cut-off values are 0.9 to signify excellent, 0.8 for good, and 0.7 for acceptable internal consistency. For some scales, some countries opted to delete one or two items. Strictly speaking, this constituted a different scale and, therefore, a footnote was added in the tables to note which item had been deleted.

### Cross-country comparability

Cross-country validity of the constructs requires a thorough and closely monitored process of translation (see Chapter 5 for a description of the translation process in PISA 2015) and standardised administration. It also makes assumptions about having measured the same construct in different national and cultural contexts. All of the indicators are based on self-reports. Such measures can suffer from various measurement errors, for instance, students are asked to report their behaviour retrospectively. Cultural differences in attitudes towards self-enhancement can influence country-level results in examinees' self-reported beliefs, behaviours and attitudes (Bempechat, Jimenez and Boulay, 2002). The literature consistently shows that response biases, such as social desirability, acquiescence and extreme response choice, are more common in countries with lower socio-economic development, compared with more affluent countries. Within countries, these response styles differ between gender and across socio-economic status levels (Buckley, 2009).

Psychometric techniques can be used to analyse the extent to which the measurement of the latent constructs is consistent across participating countries, thus indicating whether the measured construct can be compared across countries. In PISA 2015, cross-country comparability was investigated via two different approaches:

- For each scale in each country, the **internal consistency** was calculated (see above).
- For each item and scale, analyses on the **invariance of item parameters** across countries and languages within a country were conducted.

**Internal consistency.** The Cronbach's alpha coefficient of internal consistency will be reported for each country along with each scaled construct in the different questionnaire sections in this chapter. Similar and high values across countries are a good indication about having measured reliably across countries.

**Invariance of item parameters.** PISA 2015 implemented an innovative approach to test whether equal (*invariant*) item parameters can be assumed across groups of participating countries and language groups therein. In a first step, groups were defined whereas every country or multiple, sufficiently large samples of examinees taking the same questionnaire language version within the country formed one group each. For regular scales, groups are based on country-by-language combinations; for trend scales, groups are based on cycle-by-country-by-language combinations. A senate-weighted sample size of at least 300 cases was considered sufficiently large to form one group. In a second step, international item and person parameters were estimated based on all examinees across all groups (see section "Scaling procedures").

Based on this estimation, the root mean square deviance (*RMSD*) item-fit statistic was calculated for each group and item by:

**16.9**

$$RMSD = \sqrt{\int (P_o(\theta) - P_e(\theta))^2 f(\theta) d\theta}$$

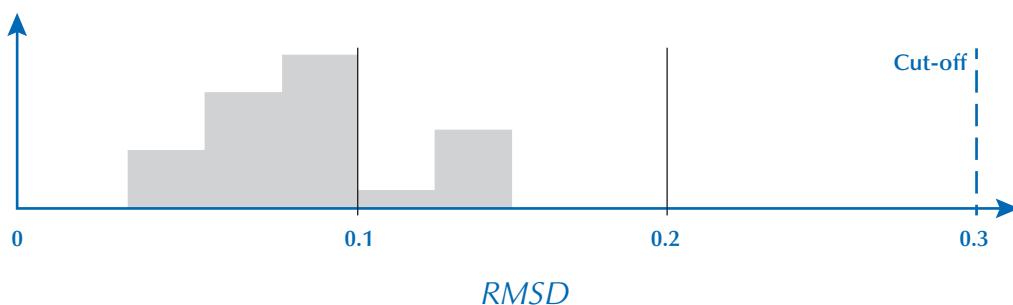
quantifying the difference between the *observed item characteristic curve*<sup>3</sup> (*ICC*,  $P_o(\theta)$ ) with the *model-based ICC* ( $P_e(\theta)$ ). The RMSD statistic is sensitive to the group-specific deviations of both the item difficulty parameters and item slope parameters from the international parameters. Values close to zero indicate good item fit, meaning that the model with international item parameters describes the responses in this group very well. A value of 0.3 was set as a cut-off criterion, with larger values indicating that the international item parameters are not appropriate for this group. Instead, a flagged group was allowed to receive group-specific (*unique*) item parameters and steps 2 and 3 were repeated until all items exhibited *RMSD* values smaller than 0.3.<sup>4</sup> The final distribution of *RMSD* values across groups will be reported for each item along with each of the scales. (For an explanation of the graphical representation, see section “*Evaluating cross-country comparability*” below.)

#### **Evaluating cross-country comparability of latent constructs**

PISA 2015 adopted a new approach to evaluating the invariance of latent constructs across groups. The *RMSD* statistic quantifies how well the international parameters describe a group’s observed data, and its distribution across groups indicates the international item parameters’ fit, i.e., how well the international item parameters function across groups. The histogram of this distribution will be referred to as *RMSD-plot* and will be reported along with each item’s wording and parameters in the subsequent sections in which each scale is presented individually. Figure 16.3 gives an example of such a plot.

■ Figure 16.3 ■

#### **Example of an RMSD-plot: distribution of the RMSD statistic across groups**

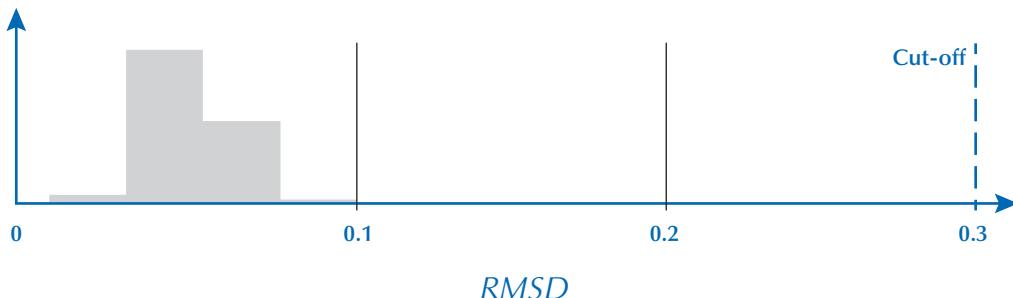


The histogram shows the distribution of *RMSD* values for a sample item across all groups. Blue elements were added for illustration but will be omitted from the plots reported along with each item in the subsequent sections. The x-axis is held constant for all plots, ranging from its theoretical minimum to maximum. The theoretical minimum (*RMSD*=0) indicates perfect fit of the international item parameters for this group. A value of *RMSD*=0.3 was chosen as the cut-off criterion for assigning group-specific parameters, thus indicating the possible maximum of *RMSD*. Vertical lines in black signify *RMSD* values of 0.1 and 0.2, respectively. Figure 16.3 would therefore indicate good item fit in most groups, with only very few groups exhibiting values larger than 0.1. Figure 16.4, in comparison, would indicate very good fit in all countries, thus signifying high cross-country comparability of the construct.



■ Figure 16.4 ■

**Example of an RMSD-distribution for a very well fitting item across all groups:  
All RMSD values are less than 0.1**



Annex H contains the complete documentation of RMSD values for each construct's items and each group.

### STUDENT QUESTIONNAIRE DERIVED VARIABLES

The Student Questionnaire features 54 derived variables, both simple questionnaire indices as well as scaled variables. Moreover, information from the Student Questionnaire was used to calculate the ESCS composite scores. The derived variables are shown in Table 16.3 and will be described in the following. Simple questionnaire indices are preceded by those that are based on IRT scaling.

[Part 1/2]

Table 16.3 Derived variables in the PISA 2015 Student Questionnaire

DV Name	Description	Question no.	Trend to PISA 2006	IRT scaling
GRADE	Grade compared to modal grade in country	ST001		
AGE	Age	ST003		
MISCED	Mother's education (ISCED)	ST005, ST006		
HISCED	Highest education of parents (ISCED)	ST005, ST006, ST007, ST008		
FISCED	Father's education (ISCED)	ST007, ST008		
PARED	Highest education of parents in years	ST005, ST006, ST007, ST008		
CULTPOSS	Cultural possessions at home	ST011, ST012		YES
HEDRES	Home educational resources	ST011		YES
WEALTH	Family wealth	ST011, ST012		YES
ICTRES	ICT Resources	ST011, ST012		YES
HOMEPOS	Home possessions	ST011, ST012, ST013		YES
ESCS	Index of economic, social and cultural status	ST005, ST006, ST007, ST008, ST011, ST012, ST013, ST014, ST015		
BFMJ2	ISEI of father	ST015		
BMMJ1	ISEI of mother	ST014		
HISEI	Highest parental occupational status	ST014, ST015		
IMMIG	Immigration status	ST019		
COBN_F	Country of birth national categories – father	ST019Q01TA		
COBN_M	Country of birth national categories – mother	ST019Q01TB		
COBN_S	Country of birth national categories – student	ST019Q01TC		
LANGN	Language at home	ST022Q01TA		
BELONG	Sense of Belonging to School	ST034		YES
beingbullied	Being Bullied	ST038		
unfairteacher	Teacher Fairness	ST039		
LMINS	Learning time (minutes per week) – <test language>	ST059, ST061		
MMINS	Learning time (minutes per week) – <mathematics>	ST059, ST061		
SMINS	Learning time (minutes per week) – <science>	ST059, ST061		
TMINS	Learning time (minutes per week) – in total	ST060, ST061		
OUTHOURS	Out-of-school study time per week	ST071		
COOPERATE	Enjoy cooperation	ST082		YES
CPSVALUE	Value cooperation	ST082		YES
ENVAWARE	Environmental awareness	ST092	YES	YES
ENVOPT	Environmental optimism	ST093	YES	YES
JOYSCIE	Enjoyment of science	ST094	YES	YES
INTBRSCI	Interest in broad science topics	ST095		YES
DISCLISCI	Disciplinary climate in science classes	ST097		YES
IBTEACH	Inquiry-based science teaching and learning practices	ST098		YES

[Part 2/2]

Table 16.3 Derived variables in the PISA 2015 Student Questionnaire

DV Name	Description	Question no.	Trend to PISA 2006	IRT scaling
TEACHSUP	Teacher support in a science classes	ST100		YES
TDTEACH	Teacher-directed science instruction	ST103		YES
PERFEED	Perceived feedback	ST104		YES
ADINST	Adaption of instruction	ST107		YES
INSTSCIE	Instrumental motivation	ST113	YES	YES
BSMJ	Students' expected occupational status (SEI)	ST114		
ANXTEST	Test anxiety	ST118		YES
MOTIVAT	Achieving motivation	ST119		YES
EMOSUPS	Parents emotional support	ST123		YES
DURECEC	Duration in early childhood education and care	ST125, ST126		
REPEAT	Grade repetition	ST127		
SCIEEFF	Science self-efficacy	ST129	YES	YES
EPIST	Epistemological beliefs	ST131		YES
SCIEACT	Science activities	ST146	YES	YES
ISCEDD	ISCED designation	ST (SPT)		
ISCEDL	ISCED level	ST (SPT)		
ISCEDO	ISCED orientation	ST (SPT)		
PROGN	Unique national study programme code	ST002		

### Grade

The relative grade index (GRADE) was computed to capture between-country variation. It indicates whether students are in the country's a modal grade i (value of 0) or whether they are below or above the modal grade (+x grades, -x grades). The information about the students' grade level was taken from the Student Questionnaire (ST001) whereas the modal grade was defined by the country and documented in the student tracking form.

### Student age

The age of a student (AGE) was calculated as the difference between the year and month of the testing and the year and month of a student's birth. Data on student's age were obtained from both the questionnaire (ST003) and the student tracking forms. If the month of testing was not known for a particular student, the median month for that country was used in the calculation. The formula for computing AGE was:

$$AGE = (100 + T_y - S_y) + (T_m - S_m)/12$$

where  $T_y$  and  $S_y$  are the year of the test and the year of the students' birth, respectively in two-digit format (for example "06" or "92"), and  $T_m$  and  $S_m$  are the month of the test and month of the students' birth, respectively. The result is rounded to two decimal places.

### Educational level of parents

Students' responses on questions ST005, ST006, ST007, and ST008 regarding parental education were classified using ISCED 1997 (OECD, 1999). Indices on parental education were constructed by recoding educational qualifications into the following categories: (0) None, (1) ISCED 1 (primary education), (2) ISCED 2 (lower secondary), (3) ISCED Level 3B or 3C (vocational/pre-vocational upper secondary), (4) ISCED 3A (general upper secondary) and/or ISCED 4 (non-tertiary post-secondary), (5) ISCED 5B (vocational tertiary) and (6) ISCED 5A and/or ISCED 6 (theoretically oriented tertiary and post-graduate). Indices with these categories were provided for a student's mother (MISCED) and father (FISCED). In addition, the index of highest educational level of parents (HISCED) corresponds to the higher ISCED level of either parent. The index of highest educational level of parents was also recoded into estimated number of years of schooling (PARED).<sup>5</sup>

### Highest occupational status of parents

Occupational data for both the student's father and student's mother were obtained from responses to open-ended questions. The responses were coded to four-digit ISCO codes and then mapped to the international socio-economic index of occupational status (ISEI) (Ganzeboom and Treiman, 2003). In PISA 2015, the new ISCO and ISEI in their 2008 version were used. Three indices were calculated based on this information: father's occupational status (BFMJ2); mother's occupational status (BMMJ1); and the highest occupational status of parents (HISEI) which corresponds to the



higher ISEI score of either parent or to the only available parent's ISEI score. For all three indices, higher ISEI scores indicate higher levels of occupational status.

### **Immigration background**

The PISA database contains three country-specific variables relating to the students' country of birth, their mother and father (COBN\_S, COBN\_M, and COBN\_F). The items ST019Q01TA, ST019Q01TB and ST019Q01TC were recoded into the following categories: (1) country of birth is the same as country of assessment and (2) other. The index of immigrant background (IMMIG) was calculated from these variables with the following categories: native students (those students who had at least one parent born in the country), (2) second-generation students (those born in the country of assessment but whose parent(s) were born in another country) and (3) first-generation students (those students born outside the country of assessment and whose parents were also born in another country). Students with missing responses for either the student or for both parents were assigned missing values for this variable.

### **Language spoken at home**

Students indicated what language they usually speak at home (ST022), and the database includes a derived variable (LANGN) containing a country-specific code for each language. In addition, an internationally comparable variable was derived from this information with the following categories: (1) language at home is the same as the language of assessment for that student and (2) language at home is another language.

### **School climate regarding fairness and bullying**

PISA 2015 included two new questions on being bullied (ST038) and teacher fairness (ST039) asking students about how often in the past 12 months they had experienced bullying behaviour of other students or unfair treatment by teachers. The questions used a four-point scale distinguishing the answer categories "never or almost never", "a few times a year", "a few times a month", "once a week or more". The derived variable TEACHFAIR reports a mean for each scale. However, as the data for ST038 showed a strongly skewed distribution, no scale was built. Results should be used with caution and cross-country comparability needs to be investigated further.

### **Learning time**

Learning time in test language (LMINS) was computed by multiplying the number of minutes on average in the test language class by number of test language class periods per week (ST061 and ST059). Comparable indices were computed for mathematics (MMINS) and science (SMINS). Learning time in total (TMINS) was computed using information about the average minutes in a <class period> (ST061) in relation to information about the number of class periods per week attended in total (ST060).

### **Out-of-school study time**

Students were asked in a slider-format question how much time they spent studying in addition to their required school schedule (ST071). The index OUTHOURS was computed by summing the time spent studying for different school subjects.

### **Expected occupational status**

As in previous cycles of PISA, students were asked to report their expected occupation at age 30 and a description of this job. The responses were coded to four-digit ISCO codes and then mapped to the ISEI index (Ganzeboom et al., 2003). Recoding of ISCO codes into ISEI index results in scores for the students' expected occupational status (BSMJ), where higher scores of ISEI indicate higher levels of expected occupational status.

### **Early childhood education and care**

Questions ST125 and ST126 measure the starting age in ISCED 1 and ISCED 0. A difference score of the two thus indicates the number of years a student spent in early childhood education and care. This indicator is called DURECEC.

### **Grade repetition**

The grade repetition variable (REPEAT) was computed by recoding variables ST127Q01TA, ST127Q02TA, and ST127Q03TA. REPEAT took the value of "1" if the student had repeated a grade in at least one ISCED level and the value of "0" if "no, never" was chosen at least once, given that none of the repeated grade categories were chosen. The index is assigned a missing value if none of the three categories were ticked in any levels.



### **Study programme indices**

PISA collects data on study programmes available to 15-year old students in each country. This information is obtained through the student tracking form and the Student Questionnaire. In the final database, all national programmes are included in a separate derived variable (PROGN) where the first six digits represent the National Centre code, and the last two digits are the nationally specific programme code. All study programmes were classified using the International Standard Classification of Education (ISCED 1997)<sup>6</sup>. The following indices were derived from the data on study programmes: programme level (ISCEDL) indicates whether students were at the lower or upper secondary level (ISCED 2 or ISCED 3); programme designation (ISCEDD) indicates the designation of the study programme (A = general programmes designed to give access to the next programme level, B = programmes designed to give access to vocational studies at the next programme level, C = programmes designed to give direct access to the labour market, M = modular programmes that combine any or all of these characteristics); and programme orientation (ISCEDO) indicates whether the programme's curricular content was general, pre-vocational or vocational.

### **Derived variables based on IRT Scaling**

The PISA 2015 Student Questionnaire provided data for 25 scaled indices which will be presented along with the item content and parameters in the following.

#### **Household possessions**

In PISA 2015, students reported the availability of 16 household items at home (ST011) including three country-specific household items that were seen as appropriate measures of family wealth within the country's context. In addition, students reported the amount of possessions and books at home (ST012, ST013). Five indices were derived from these items: i) family wealth possessions (WEALTH), ii) cultural possessions (CULTPOSS), iii) home educational resources (HEDRES), iv) ICT resources (ICTRES) and v) home possessions (HOMEPOS). Table 16.4 gives an overview of the indicator items for each of these five indices.

**Table 16.4 Indicators of household possessions and home background indices**

Item	Description	Item is used to measure index				
		HOMEPOS	WEALTH	CULTPOSS	HEDRES	ICTRES
ST011Q01TA	A desk to study at	X			X	
ST011Q02TA	A room of your own	X	X			
ST011Q03TA	A quiet place to study	X			X	
ST011Q04TA	A computer you can use for school work	X			X	
ST011Q05TA	Educational software	X			X	X
ST011Q06TA	A link to the Internet	X	X			X
ST011Q07TA	Classic literature (e.g. <Shakespeare>)	X		X		
ST011Q08TA	Books of poetry	X		X		
ST011Q09TA	Works of art (e.g. paintings)	X		X		
ST011Q10TA	Books to help with your school work	X			X	
ST011Q11TA	<Technical reference books>	X			X	
ST011Q12TA	A dictionary	X			X	
ST011Q16NA	Books on art, music, or design	X		X		
ST011Q17TA	<Country-specific wealth item 1>	X	X			
ST011Q18TA	<Country-specific wealth item 2>	X	X			
ST011Q19TA	<Country-specific wealth item 3>	X	X			
ST012Q01TA	Televisions	X	X			
ST012Q02TA	Cars	X	X			
ST012Q03TA	Rooms with a bath or shower	X	X			
ST012Q05NA	<Cell phones> with Internet access (e.g. smartphones)	X	X			X
ST012Q06NA	Computers (desktop computer, portable laptop, or notebook)	X	X			X
ST012Q07NA	<Tablet computers> (e.g. <iPad®>, <BlackBerry® PlayBook™>)	X	X			X
ST012Q08NA	E-book readers (e.g. <KindleTM>, <Kobo>, <Bookeen>)	X	X			X
ST012Q09NA	Musical instruments (e.g. guitar, piano)	X		X		
ST013Q01TA	How many books are there in your home?	X				



Tables 16.5 and 16.6 provide information on the reliabilities (Cronbach's Alpha coefficients) in OECD countries and partner countries and economies, respectively.

**Table 16.5 Scale reliabilities for Household possessions indices in OECD countries**

	HOMEPOS	CULTPOSS	HEDRES	WEALTH	ICTRES
Australia	0.734	0.575	0.647	0.640	0.481
Austria	0.728	0.586	0.507	0.664	0.478
Belgium	0.731	0.624	0.524	0.667	0.523
Canada	0.730	0.584	0.629	0.649	0.520
Chile	0.809	0.571	0.541	0.750	0.626
Czech Republic	0.715	0.626	0.550	0.628	0.480
Denmark	0.684	0.597	0.504	0.559	0.371
Estonia	0.741	0.576	0.493	0.682	0.477
Finland	0.706	0.643	0.544	0.558	0.427
France	0.712	0.657	0.496	0.634	0.487
Germany	0.714	0.601	0.522	0.624	0.501
Greece	0.752	0.581	0.498	0.699	0.562
Hungary	0.780	0.650	0.555	0.711	0.516
Iceland	0.693	0.530	0.581	0.630	0.400
Ireland	0.730	0.582	0.550	0.608	0.465
Israel <sup>1</sup>	0.737	0.634	0.587	0.696	0.545
Italy	0.732	0.557	0.491	0.651	0.523
Japan	0.698	0.588	0.472	0.565	0.524
Korea	0.779	0.631	0.552	0.627	0.482
Latvia	0.723	0.584	0.420	0.646	0.503
Luxembourg	0.761	0.610	0.556	0.698	0.526
Mexico	0.867	0.601	0.574	0.847	0.739
Netherlands	0.678	0.574	0.498	0.570	0.424
New Zealand	0.748	0.561	0.653	0.673	0.549
Norway	0.726	0.621	0.608	0.636	0.445
Poland	0.748	0.598	0.456	0.690	0.496
Portugal	0.771	0.598	0.478	0.672	0.550
Slovak Republic	0.780	0.618	0.675	0.695	0.548
Slovenia	0.720	0.620	0.472	0.634	0.477
Spain	0.755	0.598	0.510	0.656	0.555
Sweden	0.748	0.611	0.608	0.653	0.473
Switzerland	0.702	0.587	0.529	0.616	0.492
Turkey	0.855	0.641	0.650	0.773	0.673
United Kingdom	0.748	0.631	0.629	0.638	0.501
United States	0.802	0.593	0.660	0.692	0.578

1. In Israel, items ST011Q02TA and ST012Q03TA were not included.

HOMEPOS is a summary index of all household and possession items (ST011, ST012 and ST013). HOMEPOS is also one of three components in the construction of the PISA index of economic, social and cultural status (or ESCS; see the section on ESCS index construction later in this chapter). The home possessions scale for PISA 2015 was computed differently than in the previous cycles. The IRT model has changed for all cognitive and non-cognitive scales for the purpose of cross-cultural comparability (See section "Cross-country comparability" in this chapter). Categories for the number of books in the home are unchanged in PISA 2015. The ST011-items (1="yes", 2="no") were reverse-coded so that a higher level indicates the presence of the indicator. Please note that items ST011Q17- ST011Q19 represent national indicators of home possessions (see Annex E) and thus differ in meaning across countries. Item parameters were therefore allowed to vary across countries during calibration and are provided in Tables 16.7 and 16.8 for OECD countries and partner countries and economies, respectively.

**Table 16.6 Scale reliabilities for Household possessions indices in partner countries and economies**

	HOMEPOS	CULTPOSS	HEDRES	WEALTH	ICTRES
Albania	0.782	0.431	0.598	0.766	0.715
Algeria	0.811	0.572	0.689	0.744	0.662
Argentina	0.810	0.595	0.587	0.726	0.584
B-S-J-G (China)*	0.868	0.658	0.650	0.814	0.713
Brazil	0.832	0.515	0.586	0.797	0.660
Bulgaria	0.784	0.573	0.580	0.740	0.581
Colombia	0.863	0.575	0.584	0.817	0.727
Costa Rica	0.859	0.603	0.584	0.814	0.676
Croatia	<b>0.744</b>	0.623	0.463	0.656	0.470
Cyprus <sup>1</sup>	0.780	0.602	0.600	0.713	0.606
Dominican Republic	0.861	0.560	0.591	0.835	0.721
FYROM	0.775	0.570	0.558	0.689	0.574
Georgia	0.809	0.604	0.510	0.735	0.625
Hong Kong (China)	0.800	0.605	0.583	0.697	0.516
Indonesia	0.855	0.582	0.621	0.806	0.752
Jordan	0.848	0.624	0.709	0.798	0.699
Kazakhstan	0.794	0.564	0.598	0.701	0.514
Kosovo	0.774	0.498	0.522	0.713	0.611
Lebanon <sup>2</sup>	0.798	0.576	0.559	0.700	0.542
Lithuania	0.775	0.635	0.504	0.696	0.515
Macao (China)	0.787	0.596	0.570	0.714	0.484
Malaysia <sup>3</sup>	0.804	0.543	0.562	0.756	0.680
Malta	0.726	0.570	0.624	0.632	0.515
Moldova	0.823	0.566	0.609	0.779	0.681
Montenegro	0.798	0.588	0.602	0.752	0.619
Peru	0.869	0.513	0.622	0.852	0.735
Qatar	0.791	0.567	0.694	0.788	0.617
Romania	0.785	0.501	0.545	0.706	0.544
Russia	0.760	0.535	0.521	0.716	0.573
Singapore	0.795	0.627	0.614	0.704	0.558
Chinese Taipei	0.785	0.678	0.597	0.648	0.527
Thailand	0.843	0.556	0.632	0.811	0.689
Trinidad and Tobago	0.805	0.553	0.616	0.756	0.695
Tunisia	0.866	0.607	0.622	0.834	0.719
United Arab Emirates	0.795	0.592	0.636	0.791	0.593
Uruguay	0.830	0.634	0.575	0.754	0.632
Viet Nam	0.823	0.610	0.569	0.787	0.664

\* B-S-J-G (China) refers to the four PISA-participating China provinces: Beijing, Shanghai, Jiangsu and Guangdong.

1. Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

2. In Lebanon, item ST012Q07NA was not included.

3. In Malaysia, item ST012Q08NA was not included.



Table 16.7 Item parameters for national home possession indicators in OECD countries

	ST011Q17TA		ST011Q18TA		ST011Q19TA	
	beta	alpha	beta	alpha	beta	alpha
Australia	1.638	0.717	1.039	0.710	0.970	1.200
Austria	-0.136	1.341	-0.730	0.674	1.187	1.250
Belgium (Flemish)	0.606	0.864	0.890	1.159	0.844	1.801
Belgium (French)	-0.238	0.720	0.505	0.812	0.902	1.712
Canada	-0.715	1.250	0.858	0.869	-0.041	0.845
Chile	0.988	0.749	0.085	1.701	-0.654	1.198
Czech Republic	N/A		N/A		N/A	
Denmark	0.847	2.910	N/A		N/A	
Estonia	0.118	1.448	-0.083	1.901	0.582	1.616
Finland	-0.012	2.205	0.314	0.870	N/A	
France	-0.169	0.918	0.036	1.428	-0.230	1.423
Germany	N/A		1.443	-0.556	0.752	2.359
Greece	0.165	1.687	-0.027	1.118	0.828	1.410
Hungary	0.058	0.718	0.156	1.134	-0.025	2.355
Iceland	0.820	1.245	1.438	1.217	1.375	0.718
Ireland	-0.223	1.138	-0.186	0.812	-1.646	0.831
Israel	1.270	0.865	0.859	1.409	1.066	1.052
Italy	0.610	1.254	0.810	1.134	0.286	1.068
Japan	-0.444	1.934	-2.544	0.724	-0.170	1.010
Korea	-0.195	1.373	0.893	1.488	0.774	1.624
Latvia	-0.616	1.768	-0.371	1.258	2.291	0.497
Luxembourg	-0.645	0.617	0.584	1.317	-1.044	0.329
Mexico	0.235	1.089	-0.377	1.254	-0.479	1.362
Netherlands	0.752	0.783	1.117	2.825	0.204	1.102
New Zealand	0.271	0.893	1.686	0.854	1.603	1.307
Norway	0.084	1.875	-0.360	0.915	N/A	
Poland	0.463	1.802	-0.093	2.369	0.066	2.315
Portugal	-1.156	0.705	-0.393	1.267	1.027	1.201
Slovak Republic	0.053	1.290	-0.044	2.296	N/A	
Slovenia	-0.887	0.604	0.421	1.834	0.326	1.529
Spain	-0.350	1.588	0.348	0.716	1.115	0.867
Sweden	1.099	1.572	1.487	0.826	0.857	0.916
Switzerland	0.532	2.438	-0.736	0.630	0.003	1.416
Turkey	-0.121	1.107	-0.236	1.735	0.902	1.487
United Kingdom (excl. Scotland)	-0.938	0.442	-0.865	0.940	-0.282	1.389
United Kingdom (Scotland)	N/A		0.639	1.619	1.900	1.472
United States	0.901	0.871	-0.197	1.392	0.469	1.258

**Notes:**

- N/A indicates that no data on the item were available for calibration.
- Both Belgium (Flemish and French) and United Kingdom (excl. Scotland) and United Kingdom (Scotland) were treated as two separate entities each during calibration and are therefore listed twice each.

**Table 16.8 Item parameters for national home possession indicators in partner countries and economies**

	ST011Q17TA		ST011Q18TA		ST011Q19TA	
	beta	alpha	beta	alpha	beta	alpha
Albania	-0.898	1.362	-0.723	1.465	-1.174	1.004
Algeria		N/A		N/A		N/A
Argentina	0.484	0.346	0.035	1.244	-1.125	1.783
Brazil	0.443	1.403	-0.063	1.230	0.782	1.493
B-S-J-G (China)	0.450	2.111	0.018	2.436	-0.082	1.335
Bulgaria	-0.282	2.199	0.001	2.175	0.236	1.363
Colombia	-0.129	1.721		N/A	-0.877	0.915
Costa Rica	-0.759	1.278	0.034	1.592	0.757	1.225
Croatia	-0.087	0.935	0.209	1.117	0.342	1.535
Cyprus*	1.123	1.046	1.760	1.340	1.031	1.370
Dominican Republic	-0.139	1.330	0.302	1.710	-0.539	1.282
FYROM	0.956	1.212	1.419	0.979		N/A
Georgia	0.704	1.073	0.933	1.225	1.596	1.228
Hong Kong (China)	-0.090	1.255	0.660	2.111	0.672	0.796
Indonesia	0.017	1.459	-1.723	1.063	0.168	2.101
Jordan	0.219	0.916	0.017	1.281	0.117	1.650
Kazakhstan	-0.006	2.677	0.000	2.047	-0.495	1.169
Kosovo	-1.289	1.342	-0.562	1.459	0.126	1.876
Lebanon	-0.388	1.407	-0.641	1.397	-1.654	0.680
Lithuania	0.053	2.357	0.572	1.141	0.202	1.783
Macao (China)	0.651	2.051	-0.095	2.325	0.206	1.803
Malaysia	-3.237	0.737	-2.647	0.821	0.203	1.550
Malta	1.282	1.241	1.499	0.847	1.809	1.167
Moldova	-0.070	1.910		N/A		N/A
Montenegro	-0.500	2.070	-0.122	2.063	0.010	2.284
Peru	0.310	1.438	-0.641	1.875	-0.105	2.037
Qatar	0.223	0.897	0.064	1.556	0.020	1.227
Romania	-0.759	1.192	-1.342	0.788	0.068	1.899
Russia	1.183	2.220	0.714	1.667	0.701	1.464
Singapore	-0.049	1.836	1.109	1.471		N/A
Chinese Taipei	0.560	2.208	-0.232	1.333	-0.187	1.744
Thailand	0.238	2.405	1.625	1.304	0.180	1.981
Trinidad and Tobago	-0.559	1.159	-1.388	0.813	-0.793	0.536
Tunisia	-0.410	1.679	-0.068	1.912	-1.374	1.295
United Arab Emirates	0.185	1.389	0.205	1.297	0.157	1.432
Uruguay	-0.088	0.487	-1.176	1.585	-0.072	2.496
Viet Nam	0.087	2.643	-2.239	0.989	0.566	2.121

\* See note under Table 16.6.

Note: N/A indicates that no data on the item were available for calibration.

Tables 16.9, 16.10, 16.11, 16.12 and 16.13 show the item wording, international item parameters and item fit for each of the five scales, respectively. Please note that all items of question ST011 are dichotomous, resulting in a 2PL model with only two item parameters: one referring to item difficulty ( $\beta$ ) and one referring to item discrimination ( $\alpha$ ). No threshold parameters ( $d$ ) are necessary.

**Table 16.9 Item parameters for Home possessions (HOMEPOS)**

Item	Description	Parameter estimates						
		beta	d_1	d_2	d_3	d_4	d_5	alpha
<b>ST011</b>	<b>Which of the following are in your home?</b>							
ST011Q01TA	A desk to study at	-0.99622						0.99603
ST011Q02TA	A room of your own	-0.81525						0.76710
ST011Q03TA	A quiet place to study	-1.13652						0.81346
ST011Q04TA	A computer you can use for school work	-0.34469						2.02990
ST011Q05TA	Educational software <sup>1</sup>	0.34028						0.95189
ST011Q06TA	A link to the Internet	-0.41684						2.44836
ST011Q07TA	Classic literature (e.g. <Shakespeare>)	*						*
ST011Q08TA	Books of poetry	*						*
ST011Q10TA	Books to help with your school work <sup>2</sup>	-1.22602						0.59293
ST011Q11TA	<Technical reference books>	0.18772						0.88643
ST011Q12TA	A dictionary	-1.74582						0.70110
ST011Q16NA	Books on art, music, or design <sup>3</sup>	-1.02696						1.25556
ST011Q17TA	<Country-specific wealth item 1>	*						*
ST011Q18TA	<Country-specific wealth item 2>	*						*
ST011Q19TA	<Country-specific wealth item 3>	*						*
<b>ST012</b>	<b>How many of these are there at your home?</b>							
ST012Q01TA	Televisions	-0.73991	1.90507	-0.71847	-1.18659			0.62294
ST012Q02TA	Cars	0.56249	0.74369	-0.05607	-0.68762			0.97934
ST012Q03TA	Rooms with a bath or shower	0.43739	1.35552	-0.41649	-0.93904			0.98154
ST012Q05NA	<Cell phones> with Internet access (e.g. smartphones)	-0.45208	0.36189	-0.50701	0.14512			0.83810
ST012Q06NA	Computers (desktop computer, portable laptop, or notebook)	0.20563	0.63235	-0.16855	-0.46379			1.69130
ST012Q07NA	<Tablet computers> (e.g. <iPad®>, <BlackBerry® PlayBook™>)	0.81206	0.48676	-0.30489	-0.18187			0.87564
ST012Q08NA	E-book readers (e.g. <Kindle™>, <Kobo>, <Bookeen>)	1.79575	-0.24104	-0.25426	0.49529			0.64692
ST012Q09NA	Musical instruments (e.g. guitar, piano)	0.88257	0.12460	-0.30754	0.18294			0.65086
<b>ST013Q01TA</b>	<b>How many books are there in your home?</b>	<b>0.84015</b>	<b>0.67861</b>	<b>0.82937</b>	<b>-0.54141</b>	<b>-0.28625</b>	<b>-0.68033</b>	<b>0.49389</b>

\* All groups received group-specific (unique) item parameters.

1. For item ST011Q05TA, group-specific (unique) item parameters were assigned for Japan: beta = 1.08454 and alpha = 1.76169.

2. For item ST011Q10TA, group-specific (unique) item parameters were assigned for Puerto Rico: beta = 0.27360 and alpha = 1.09664.

3. For item ST011Q16NA, group-specific (unique) item parameters were assigned for Albania: beta = -1.02696 and alpha = 1.2555.

**Table 16.10 Item parameters for Family wealth (WEALTH)**

Item		Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST011Q02TA	A room of your own	-1.18794				0.59067
ST011Q06TA	A link to the Internet	-0.64913				1.85772
ST012Q01TA	Televisions	-0.79254	1.98967	-0.70424	-1.28543	0.65754
ST012Q02TA	Cars	0.56392	0.85004	-0.06279	-0.78725	0.99954
ST012Q03TA	Rooms with a bath or shower	0.42688	1.57095	-0.46515	-1.10579	0.89156
ST012Q06NA	Computers (desktop computer, portable laptop, or notebook)	0.13353	0.77663	-0.21262	-0.56402	1.32688
ST012Q07NA	<Tablet computers> (e.g. <iPad®>, <BlackBerry® PlayBook™>)	0.84143	0.57964	-0.29950	-0.28014	0.91206
ST012Q08NA	E-book readers (e.g. <Kindle™>, <Kobo>, <Bookeen>)	2.19905	-0.35458	-0.32916	0.68374	0.48155

**Table 16.11 Item parameters for Cultural possessions at home (CULTPOSS)**

Item		Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST011Q07TA	Classic literature (e.g. <Shakespeare>)	-0.08572				1.48509
ST011Q08TA	Books of poetry	-0.01282				1.61409
ST011Q09TA	Works of art (e.g. paintings)	-0.42053				0.73223
ST011Q16NA	Books on art, music, or design	-0.24687				0.92627
ST012Q09NA	Musical instruments (e.g. guitar, piano)	0.94172	-0.03097	-0.77936	0.81034	0.24232

**Table 16.12 Item parameters for Home educational resources (HEDRES)**

Item		Parameter estimates				
		beta		alpha		
ST011Q01TA	A desk to study at	-0.38085				1.09535
ST011Q03TA	A quiet place to study	-0.53925				0.84215
ST011Q04TA	A computer you can use for school work	0.09232				1.74465
ST011Q05TA	Educational software	1.03471				1.03415
ST011Q10TA	Books to help with your school work	-0.36705				0.71414
ST011Q11TA	<Technical reference books>	0.84302				0.87760
ST011Q12TA	A dictionary	-1.21037				0.69196

**Table 16.13 Item parameters for ICT Resources (ICTRES)**

Item		Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST011Q05TA	Educational software <sup>1</sup>	0.02534				0.60517
ST011Q06TA	A link to the Internet	-0.95801				1.88514
ST012Q05NA	<Cell phones> with Internet access (e.g. smartphones)	-0.96009	0.52775	-0.61056	0.08281	0.70661
ST012Q06NA	Computers (desktop computer, portable laptop, or notebook)	-0.11449	0.82147	-0.19140	-0.63006	1.56852
ST012Q07NA	<Tablet computers> (e.g. <iPad®>, <BlackBerry® PlayBook™>)	0.64927	0.64903	-0.33429	-0.31474	0.80477
ST012Q08NA	E-book readers (e.g. <Kindle™>, <Kobo>, <Bookeen>)	2.16928	-0.39790	-0.37063	0.76852	0.42979

1. For item ST011Q05TA, group-specific (unique) item parameters were assigned for Japan: beta=1.12478 and alpha=1.76169.

### Sense of belonging

PISA 2015 asked students about their sense of belonging to school (ST034) using six trend items previously used in PISA 2012 (ID in 2012: ST87). The answering format was a four-point Likert scale with the answering categories “strongly agree”, “agree”, “disagree”, and “strongly disagree”; the derived IRT scale is named BELONG. Items ST034Q02TA, ST034Q03TA and ST034Q05TA were reverse-coded so that higher WLEs and higher difficulty correspond to higher level of sense of belonging on all items.

Tables 16.14 and 16.15 contain the scale’s reliabilities (Cronbach’s Alpha) across all participating OECD and partner countries and economies, respectively.

**Table 16.14 Scale reliabilities for BELONG in OECD countries**

	BELONG
Australia	0.856
Austria	0.881
Belgium	0.795
Canada	0.850
Chile	0.839
Czech Republic	0.802
Denmark	0.862
Estonia	0.826
Finland	0.863
France	0.709
Germany	0.853
Greece	0.825
Hungary	0.848
Iceland	0.902
Ireland	0.858
Israel	N/A
Italy	0.812
Japan	0.809
Korea	0.795
Latvia	0.842
Luxembourg	0.823
Mexico	0.872
Netherlands	0.846
New Zealand	0.831
Norway	0.861
Poland	0.836
Portugal	0.830
Slovak Republic	0.808
Slovenia	0.847
Spain	0.876
Sweden	0.897
Switzerland	0.826
Turkey	0.851
United Kingdom	0.843
United States	0.857

Note: N/A indicates that the question has not been administered in the country.

**Table 16.15 Scale reliabilities for BELONG in partner countries and economies**

	BELONG
Albania	0.602
Algeria	0.649
Argentina	0.687
B-S-J-G (China)	0.792
Brazil	0.832
Bulgaria	0.801
Colombia	0.849
Costa Rica	0.891
Croatia	0.860
Cyprus*	0.828
Dominican Republic	0.858
FYROM	0.689
Georgia	0.665
Hong Kong (China)	0.782
Indonesia	0.597
Jordan	0.656
Kazakhstan	0.721
Kosovo	0.562
Lebanon	0.610
Lithuania	0.817
Macao (China)	0.762
Malaysia	0.759
Malta	0.768
Moldova	0.704
Montenegro	0.781
Peru	0.767
Qatar	0.776
Romania	0.695
Russia	0.834
Singapore	0.841
Chinese Taipei	0.867
Thailand	0.713
Trinidad and Tobago	0.741
Tunisia	0.579
United Arab Emirates	0.697
Uruguay	0.857
Viet Nam	0.612

\* See note under Table 16.6.



Table 16.16 shows the item wording, international item parameters and item fit for BELONG.

**Table 16.16 Item parameters for Sense of Belonging to School (BELONG)**

Item	Thinking about your school: to what extent do you agree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST034Q01TA	I feel like an outsider (or left out of things) at school.	-0.00458	0.56688	0.37422	-0.94110	1.21518
ST034Q02TA	I make friends easily at school.	0.00475	1.02240	0.57396	-1.59636	0.77746
ST034Q03TA	I feel like I belong at school.	0.15553	1.14692	0.59957	-1.74650	0.61414
ST034Q04TA	I feel awkward and out of place in my school.	-0.00104	0.74923	0.34099	-1.09022	1.12698
ST034Q05TA	Other students seem to like me.	0.04790	1.35674	0.85709	-2.21383	0.66787
ST034Q06TA	I feel lonely at school.	-0.07787	0.53076	0.30405	-0.83481	1.59837

### **Students' dispositions for collaborative problem solving**

PISA 2015 included a question on students' collaboration and teamwork dispositions relating to the test domain of collaborative problem solving (ST082). It asks students about their agreement to specific cooperative aspects on a four-point Likert scale with the answering categories "strongly agree", "agree", "disagree", and "strongly disagree". The question was used to build two scales, one on the enjoyment of co-operation (COOPERATE) including answers to items ST082Q02NA, ST082Q03NA, ST082Q08NA, and ST082Q12NA, and one on the value of co-operation (CPSVALUE) including answers to items ST082Q01NA, ST082Q09NA, ST082Q13NA and ST082Q14NA.

Tables 16.17 and 16.18 contain the two scales' reliabilities (Cronbach's Alpha) across all participating OECD and partner countries and economies, respectively.

**Table 16.17 Scale reliabilities for COOPERATE and CPSVALUE in OECD countries**

	COOPERATE	CPSVALUE
Australia	0.709	0.819
Austria	0.643	0.784
Belgium	0.652	0.783
Canada	0.746	0.830
Chile	0.690	0.754
Czech Republic	0.684	0.783
Denmark	0.654	0.792
Estonia	0.680	0.759
Finland	0.686	0.783
France	0.680	0.819
Germany	0.655	0.743
Greece	0.672	0.790
Hungary	0.675	0.821
Iceland	0.709	0.811
Ireland	0.671	0.833
Israel	0.726	0.754
Italy	0.607	0.791
Japan	0.683	0.794
Korea	0.700	0.822
Latvia	0.668	0.805
Luxembourg	0.695	0.821
Mexico	0.717	0.756
Netherlands	0.629	0.760
New Zealand	0.722	0.817
Norway	0.728	0.826
Poland	0.626	0.811
Portugal	0.706	0.790
Slovak Republic	0.696	0.798
Slovenia	0.661	0.767
Spain	0.685	0.753
Sweden	0.731	0.784
Switzerland	0.674	0.756
Turkey	0.698	0.565
United Kingdom	0.723	0.821
United States	0.728	0.835

**Table 16.18 Scale reliabilities for COOPERATE and CPSVALUE in partner countries and economies**

	COOPERATE	CPSVALUE
Albania	N/A	N/A
Algeria	N/A	N/A
Argentina	N/A	N/A
B-S-J-G (China)	0.677	0.821
Brazil	0.667	0.692
Bulgaria	0.715	0.818
Colombia	0.618	0.659
Costa Rica	0.675	0.729
Croatia	0.702	0.784
Cyprus*	0.727	0.796
Dominican Republic	0.780	0.753
FYROM	N/A	N/A
Georgia	N/A	N/A
Hong Kong (China)	0.736	0.871
Indonesia	N/A	N/A
Jordan	N/A	N/A
Kazakhstan	N/A	N/A
Kosovo	N/A	N/A
Lebanon	N/A	N/A
Lithuania	0.705	0.824
Macao (China)	0.605	0.724
Malaysia	0.578	0.767
Malta	N/A	N/A
Moldova	N/A	N/A
Montenegro	0.699	0.753
Peru	0.656	0.699
Qatar	0.730	0.738
Romania	N/A	N/A
Russia	0.692	0.795
Singapore	0.688	0.822
Chinese Taipei	0.714	0.863
Thailand	0.648	0.716
Trinidad and Tobago	N/A	N/A
Tunisia	0.593	0.787
United Arab Emirates	0.714	0.747
Uruguay	0.657	0.756
Viet Nam	N/A	N/A

\* See note under Table 16.6.

Tables 16.19 and 16.20 show the actual item content, the international item parameters and item fit for each of the two scales, respectively.

**Table 16.19 Item parameters for Enjoy co-operation (COOPERATE)**

Item	To what extent do you disagree or agree with the following statements about yourself?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST082Q02NA	I am a good listener.	-0.15973	1.28074	0.69911	-1.97985	0.78526
ST082Q03NA	I enjoy seeing my classmates be successful.	0.00652	0.91051	0.68885	-1.59936	1.10539
ST082Q08NA	I take into account what others are interested in.	0.17180	1.12003	0.53218	-1.65221	1.27455
ST082Q12NA	I enjoy considering different perspectives.	-0.12068	1.20917	0.69511	-1.90428	0.83480

**Table 16.20 Item parameters for Value co-operation (CPSVALUE)**

Item	To what extent do you disagree or agree with the following statements about yourself?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST082Q01NA	I prefer working as part of a team to working alone.	0.26040	1.38266	0.42034	-1.80300	0.68975
ST082Q09NA	I find that teams make better decisions than individuals.	-0.04081	1.41758	0.28260	-1.70018	0.87040
ST082Q13NA	I find that teamwork raises my own efficiency.	0.15187	1.32041	0.21231	-1.53272	1.36366
ST082Q14NA	I enjoy cooperating with peers.	-0.32633	1.06557	0.60140	-1.66697	1.07619



### **Environmental awareness and optimism**

PISA 2015 took up two trend questions from PISA 2006 (ID in 2006: ST22, ST24) on students' awareness of environmental matters (ENVAWARE, ST092) and their perception of environmental issues as a concern (ENVOPT, ST093). To harmonise items across the two questions, new items were added focusing on the topics of air pollution, extinction of plants and animals and water shortage for ST092, and the increase of greenhouse gases in the atmosphere and the use of genetically modified organisms for ST093. In ST092, students rated their knowledge on a four-point scale in the following categories: "I have never heard of this", "I have heard about this but I would not be able to explain what it is really about", "I know something about this and could explain the general issue", "I am familiar with this and I would be able to explain this well". For ST093, students answered on a three-point scale with the following categories: "improve", "stay about the same", and "get worse". Therefore, the ST093-items were reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of environmental optimism. The derived variables ENVAWARE and ENVOPT were scaled using the IRT scaling model described above, allowing for a trend comparison between PISA 2006 and PISA 2015.

Tables 16.21 and 16.22 contain the two scales' reliabilities (Cronbach's Alpha) across all participating OECD and partner countries and economies, respectively.

**Table 16.21 Scale reliabilities for ENVAWARE and ENVOPT in OECD countries**

	ENVAWARE	ENVOPT
Australia	0.876	0.859
Austria	0.873	0.814
Belgium	0.862	0.861
Canada	0.877	0.874
Chile	0.862	0.899
Czech Republic	0.856	0.845
Denmark	0.854	0.767
Estonia	0.846	0.835
Finland	0.852	0.807
France	0.883	0.837
Germany	0.860	0.774
Greece	0.821	0.855
Hungary	0.854	0.872
Iceland	0.890	0.859
Ireland	0.849	0.810
Israel	0.882	0.873
Italy	0.848	0.830
Japan	0.887	0.808
Korea	0.890	0.864
Latvia	0.821	0.823
Luxembourg	0.876	0.846
Mexico	0.880	0.919
Netherlands	0.847	0.808
New Zealand	0.877	0.870
Norway	0.880	0.867
Poland	0.868	0.826
Portugal	0.894	0.904
Slovak Republic	0.875	0.894
Slovenia	0.875	0.849
Spain	0.858	0.840
Sweden	0.877	0.853
Switzerland	0.843	0.808
Turkey	0.902	0.933
United Kingdom	0.879	0.849
United States	0.871	0.865

**Table 16.22 Scale reliabilities for ENVAWARE and ENVOPT in partner countries and economies**

	ENVAWARE	ENVOPT
Albania	0.821	N/A
Algeria	0.780	N/A
Argentina	0.836	N/A
B-S-J-G (China)	0.860	0.880
Brazil	0.898	0.939
Bulgaria	0.904	0.914
Colombia	0.821	0.899
Costa Rica	0.877	0.911
Croatia	0.874	0.882
Cyprus*	0.856	0.905
Dominican Republic	0.878	0.927
FYROM	0.861	N/A
Georgia	0.844	N/A
Hong Kong (China)	0.868	0.876
Indonesia	0.830	N/A
Jordan	0.857	N/A
Kazakhstan	0.863	N/A
Kosovo	0.805	N/A
Lebanon	0.759	N/A
Lithuania	0.882	0.863
Macao (China)	0.846	0.838
Malaysia	0.873	0.877
Malta	0.860	N/A
Moldova	0.821	N/A
Montenegro	0.902	0.920
Peru	0.854	0.913
Qatar	0.889	0.895
Romania	0.768	N/A
Russia	0.879	0.892
Singapore	0.858	0.846
Chinese Taipei	0.903	0.863
Thailand	0.878	0.899
Trinidad and Tobago	0.826	N/A
Tunisia	0.810	0.858
United Arab Emirates	0.875	0.883
Uruguay	0.874	0.901
Viet Nam	0.749	N/A

\* See note under Table 16.6..

Note: N/A indicates that the question has not been administered in the country.

Tables 16.23 and 16.24 show the actual item content, the international item parameters and item fit for each of the two scales, respectively.

**Table 16.23 Item parameters for Environmental Awareness (ENVAWARE)**

Item	How informed are you about the following environmental issues?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST092Q01TA	The increase of greenhouse gases in the atmosphere	0.28250	1.21751	0.03283	-1.25034	0.75505
ST092Q02TA	The use of genetically modified organisms (<GMO>)	0.92331	1.57194	-0.11008	-1.46185	0.50088
ST092Q04TA	Nuclear waste	0.41005	1.56468	-0.08269	-1.48198	0.74670
ST092Q05TA	The consequences of clearing forests for other land use	-0.15483	0.96316	0.09813	-1.06129	0.94178
ST092Q06NA	Air pollution	-0.34475	0.91931	0.14903	-1.06834	1.57386
ST092Q08NA	Extinction of plants and animals	-0.25612	1.05537	0.08030	-1.13567	1.47363
ST092Q09NA	Water shortage	-0.14049	1.05455	0.10982	-1.16437	0.93072

**Table 16.24 Item parameters for Environmental optimism (ENVOPT)**

Item	Do you think problems associated with the environmental issues below will improve or get worse over the next 20 years?	Parameter estimates			
		beta	d_1	d_2	alpha
ST093Q01TA	Air pollution	0.08759	0.05125	-0.05125	1.07684
ST093Q03TA	Extinction of plants and animals	0.06571	0.34506	-0.34506	1.16385
ST093Q04TA	Clearing of forests for other land use	0.13378	0.26068	-0.26068	1.17143
ST093Q05TA	Water shortages	-0.11964	0.40556	-0.40556	1.05629
ST093Q06TA	Nuclear waste	0.04693	0.46062	-0.46062	0.84528
ST093Q07NA	The increase of greenhouse gases in the atmosphere	0.10669	0.33837	-0.33837	1.21447
ST093Q08NA	The use of genetically modified organisms (<GMO>)	-0.25762	0.64808	-0.64808	0.66175

### Interest in science

Interest in science was assessed with two scales, students' enjoyment of science (ST094) and their interest in broad science topics (ST095). Tables 16.25 and 16.26 contain the two scales' reliabilities (Cronbach's Alpha) across all participating OECD and partner countries and economies, respectively.



Table 16.25 Scale reliabilities for JOYSCIE and INTBRSCI in OECD countries

	JOYSCIE	INTBRSCI
Australia	0.956	0.826
Austria	0.945	0.766
Belgium	0.935	0.813
Canada	0.948	0.791
Chile	0.935	0.832
Czech Republic	0.914	0.796
Denmark	0.960	0.814
Estonia	0.930	0.756
Finland	0.945	0.831
France	0.924	0.802
Germany	0.945	0.765
Greece	0.934	0.799
Hungary	0.935	0.782
Iceland	0.970	0.894
Ireland	0.948	0.802
Israel	0.950	0.844
Italy	0.926	0.771
Japan	0.947	0.807
Korea	0.959	0.826
Latvia	0.919	0.719
Luxembourg	0.941	0.815
Mexico	0.899	0.826
Netherlands	0.953	0.820
New Zealand	0.945	0.808
Norway	0.963	0.855
Poland	0.919	0.763
Portugal	0.928	0.830
Slovak Republic	0.919	0.825
Slovenia	0.933	0.771
Spain	0.935	0.775
Sweden	0.968	0.852
Switzerland	0.934	0.766
Turkey	0.945	0.852
United Kingdom	0.949	0.821
United States	0.946	0.808

Table 16.26 Scale reliabilities for JOYSCIE and INTBRSCI in partner countries and economies

	JOYSCIE	INTBRSCI
Albania	0.883	N/A
Algeria	0.795	N/A
Argentina	0.881	N/A
B-S-J-G (China)	0.940	0.787
Brazil	0.911	0.850
Bulgaria	0.924	0.836
Colombia	0.903	0.826
Costa Rica	0.921	0.807
Croatia	0.940	0.806
Cyprus*	0.936	0.846
Dominican Republic	0.923	0.873
FYROM	0.898	N/A
Georgia	0.904	N/A
Hong Kong (China)	0.953	0.816
Indonesia	0.857	N/A
Jordan	0.884	N/A
Kazakhstan	0.912	N/A
Kosovo	0.919	N/A
Lebanon	0.823	N/A
Lithuania	0.933	0.769
Macao (China)	0.933	0.754
Malaysia	0.930	0.809
Malta	0.936	N/A
Moldova	0.818	N/A
Montenegro	0.938	0.840
Peru	0.914	0.822
Qatar	0.936	0.810
Romania	0.787	N/A
Russia	0.922	0.817
Singapore	0.956	0.765
Chinese Taipei	0.953	0.797
Thailand	0.898	0.767
Trinidad and Tobago	0.916	N/A
Tunisia	0.853	0.780
United Arab Emirates	0.929	0.794
Uruguay	0.930	0.813
Viet Nam	0.869	N/A

\* See note under Table 16.6.

Note: N/A indicates that the question has not been administered in the country.

Enjoyment of science (ST094) is a trend question from PISA 2006 (ID in 2006: ST16), asking students to respond on a four-point Likert scale with the categories “strongly agree”, “agree”, “disagree”, and “strongly disagree”. The derived variable JOYSCIE was scaled using the IRT scaling model described above enabling a trend comparison between PISA 2006 and PISA 2015 at the country level. Table 16.27 shows the actual item content, the international item parameters and item fit for JOYSCIE.

**Table 16.27 Item parameters for Enjoyment of science (JOYSCIE)**

Item	How much do you disagree or agree with the statements about yourself below?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST094Q01NA	I generally have fun when I am learning <broad science> topics.	-0.03733	1.99379	0.44600	-2.43980	0.89314
ST094Q02NA	I like reading about <broad science>.	0.24044	2.18913	0.19370	-2.38282	0.96880
ST094Q03NA	I am happy working on <broad science> topics.	0.40009	2.32626	0.14198	-2.46825	0.83468
ST094Q04NA	I enjoy acquiring new knowledge in <broad science>.	-0.29106	1.89703	0.38915	-2.28617	1.14639
ST094Q05NA	I am interested in learning about <broad science>.	-0.17276	1.90115	0.29334	-2.19449	1.15698

A new question to assess students' interest in science topics was developed for PISA 2015 (ST095) including topics like the biosphere, motion and forces, energy and its transformation, the Universe and its history as well as how science can help prevent disease. Students declared their interest on a five-point Likert scale with the categories “not interested”, “hardly interested”, “interested”, “highly interested”, and “I don't know what this is”. The last category was recoded as a missing. The derived variable INTBRSCI was scaled using the IRT scaling model described above. Table 16.28 shows the actual item content, the international item parameters and item fit for INTBRSCI.

**Table 16.28 Item parameters for Interest in broad science topics (INTBRSCI)**

Item	To what extent are you interested in the following <broad science> topics?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST095Q04NA	Biosphere (e.g. ecosystem services, sustainability)	0.34847	1.01950	0.33721	-1.35670	0.69433
ST095Q07NA	Motion and forces (e.g. velocity, friction, magnetic and gravitational forces)	0.14145	0.88014	0.10062	-0.98076	1.41783
ST095Q08NA	Energy and its transformation (e.g. conservation, chemical reactions)	0.08373	0.84341	0.09190	-0.93531	1.86518
ST095Q13NA	The Universe and its history	-0.58932	0.52303	0.38720	-0.91023	0.49305
ST095Q15NA	How science can help us prevent disease	-0.58180	0.66498	0.45903	-1.12401	0.52962

### Science learning in school

PISA 2015 focused on science learning in school by including several questions about the learning environment in the science classroom. They asked how often specific activities happened in the school science course.

The questions included the disciplinary climate in science classes (DISCLISCI , ST097), enquiry-based science teaching and learning practices (IBTEACH, ST098), teacher support in a science classes (TEACHSUP, ST100), teacher-directed science instruction (TDTEACH , ST103), perceived feedback (PERFEED, ST104), adaption of instruction (ADINST, ST107) and instrumental motivation (INSTSCIE, ST113). All of these derived variables were scaled using the IRT scaling model described above.

Tables 16.29 and 16.30 contain the seven scales' reliabilities (Cronbach's Alpha) across all participating OECD and partner countries and economies, respectively.



Table 16.29 Scale reliabilities for all seven indices relating to Science learning in school in OECD countries

	DISCLSCI	IBTEACH	TEACHSUP	TDTEACH	PERFEED	ADINST	INSTSCIE
Australia	0.919	0.854	0.927	0.863	0.940	0.824	0.949
Austria	0.897	0.868	0.865	0.806	0.905	0.800	0.929
Belgium	0.899	0.842	0.888	0.773	0.881	0.725	0.917
Canada	0.902	0.872	0.922	0.866	0.928	0.840	0.937
Chile	0.891	0.874	0.905	0.791	0.917	0.796	0.928
Czech Republic	0.899	0.845	0.864	0.799	0.888	0.822	0.922
Denmark	0.883	0.828	0.881	0.810	0.911	0.781	0.925
Estonia	0.903	0.840	0.895	0.805	0.906	0.770	0.876
Finland	0.906	0.832	0.905	0.841	0.929	0.815	0.933
France	0.891	0.838	0.892	0.827	0.891	0.764	0.924
Germany	0.881	0.853	0.885	0.786	0.902	0.785	0.924
Greece	0.815	0.865	0.888	0.838	0.901	0.796	0.891
Hungary	0.911	0.845	0.893	0.818	0.889	0.804	0.908
Iceland	0.899	0.892	0.919	0.846	0.941	0.842	0.955
Ireland	0.906	0.836	0.903	0.816	0.918	0.793	0.926
Israel	0.918	0.890	0.907	0.845	0.929	0.799	0.921
Italy	0.869	0.847	0.887	0.712	0.871	0.762	0.893
Japan	0.876	0.862	0.891	0.719	0.888	0.728	0.924
Korea	0.892	0.899	0.914	0.834	0.943	0.841	0.950
Latvia	0.892	0.825	0.871	0.793	0.896	0.705	0.887
Luxembourg	0.907	0.868	0.882	0.840	0.923	0.781	0.925
Mexico	0.833	0.872	0.895	0.802	0.921	0.806	0.915
Netherlands	0.875	0.864	0.868	0.702	0.909	0.771	0.948
New Zealand	0.918	0.866	0.920	0.859	0.934	0.826	0.943
Norway	0.899	0.877	0.922	0.834	0.940	0.809	0.929
Poland	0.894	0.873	0.910	0.835	0.903	0.812	0.913
Portugal	0.911	0.885	0.930	0.887	0.941	0.876	0.958
Slovak Republic	0.898	0.872	0.885	0.817	0.893	0.784	0.899
Slovenia	0.905	0.881	0.875	0.850	0.923	0.825	0.911
Spain	0.892	0.848	0.906	0.729	0.910	0.808	0.937
Sweden	0.898	0.896	0.930	0.877	0.943	0.855	0.923
Switzerland	0.888	0.848	0.871	0.825	0.913	0.767	0.924
Turkey	0.892	0.893	0.915	0.800	0.911	0.814	0.902
United Kingdom	0.919	0.856	0.918	0.835	0.933	0.838	0.933
United States	0.904	0.890	0.918	0.872	0.944	0.833	0.925

Table 16.30 Scale reliabilities for all seven indices relating to Science learning in school in partner countries and economies

	DISCLSCI	IBTEACH	TEACHSUP	TDTEACH	PERFEED	ADINST	INSTSCIE
Albania	0.804	0.756	0.782	0.648	0.865	N/A	0.822
Algeria	0.746	0.763	0.788	0.790	0.753	N/A	0.795
Argentina	0.823	0.824	0.856	0.763	0.859	N/A	0.868
B-S-J-G (China)	0.890	0.898	0.880	0.858	0.913	0.781	0.901
Brazil	0.884	0.870	0.902	0.842	0.886	0.793	0.889
Bulgaria	0.890	0.892	0.884	0.873	0.912	0.824	0.895
Colombia	0.821	0.839	0.877	0.743	0.900	0.720	0.885
Costa Rica	0.842	0.853	0.890	0.759	0.921	0.791	0.923
Croatia	0.890	0.881	0.881	0.851	0.918	0.814	0.921
Cyprus*	0.853	0.879	0.901	0.880	0.914	0.810	0.897
Dominican Republic	0.834	0.839	0.875	0.827	0.890	0.758	0.920
FYROM	0.828	0.831	0.843	0.784	0.857	N/A	0.845
Georgia	0.819	0.813	0.802	0.746	0.863	N/A	0.851
Hong Kong (China)	0.925	0.906	0.928	0.847	0.941	0.844	0.951
Indonesia	0.775	0.769	0.684	0.690	0.791	N/A	0.876
Jordan	0.826	0.857	0.879	0.851	0.854	N/A	0.830
Kazakhstan	0.778	0.816	0.788	0.822	0.864	N/A	0.916
Kosovo	0.784	0.772	0.756	0.840	0.823	N/A	0.857
Lebanon	0.773	0.762	0.780	0.758	0.835	N/A	0.756
Lithuania	0.925	0.861	0.900	0.871	0.928	0.773	0.900
Macao (China)	0.856	0.842	0.897	0.825	0.904	0.740	0.900
Malaysia	0.849	0.837	0.877	0.854	0.905	0.782	0.903
Malta	0.886	0.819	0.910	0.791	0.904	N/A	0.923
Moldova	0.777	0.738	0.780	0.757	0.821	N/A	0.852
Montenegro	0.888	0.920	0.928	0.884	0.925	0.830	0.898
Peru	0.841	0.867	0.879	0.834	0.877	0.736	0.878
Qatar	0.897	0.903	0.908	0.882	0.915	0.810	0.895
Romania	0.776	0.764	0.785	0.563	0.743	N/A	0.826
Russia	0.906	0.882	0.884	0.839	0.905	0.769	0.895
Singapore	0.889	0.865	0.914	0.851	0.933	0.828	0.906
Chinese Taipei	0.912	0.902	0.914	0.874	0.931	0.837	0.944
Thailand	0.847	0.897	0.908	0.894	0.882	0.813	0.852
Trinidad and Tobago	0.839	0.807	0.895	0.815	0.903	N/A	0.905
Tunisia	0.800	0.860	0.877	0.841	0.846	0.723	0.840
United Arab Emirates	0.885	0.896	0.909	0.864	0.917	0.816	0.899
Uruguay	0.889	0.869	0.910	0.779	0.903	0.776	0.908
Viet Nam	0.683	0.778	0.730	0.719	0.756	N/A	0.796

\* See note under Table 16.6.

Note: N/A indicates that the question has not been administered in the country.

For ST097, students responded on a four-point Likert scale with the categories “every lesson”, “most lessons”, “some lessons” and “never or hardly ever”. Table 16.31 shows the item wording, international item parameters and item fit for DISCLISCI.

**Table 16.31 Item parameters for Disciplinary climate in science classes (DISCLISCI)**

Item	To what extent are you interested in the following <broad science> topics?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST097Q01TA	Students don't listen to what the teacher says.	0.19029	1.25309	0.51737	-1.77046	0.94803
ST097Q02TA	There is noise and disorder.	0.19407	1.22680	0.34986	-1.57666	1.29726
ST097Q03TA	The teacher has to wait a long time for students to quiet down.	-0.00888	1.07093	0.31662	-1.38755	1.14809
ST097Q04TA	Students cannot work well.	-0.33810	1.08205	0.48490	-1.56696	0.79547
ST097Q05TA	Students don't start working for a long time after the lesson begins.	-0.18866	0.99587	0.37880	-1.37468	0.81114

For ST098, students responded on a four-point Likert scale with the categories “in all lessons”, “in most lessons”, “in some lessons”, “never or hardly ever”. Therefore, the ST098-items were reverse-coded so that higher WLEs and higher difficulty correspond to higher levels enquiry-based science teaching and learning practices. Table 16.32 shows the item wording, international item parameters and item fit for IBTEACH.

**Table 16.32 Item parameters for Inquiry-based science teaching and learning practices (IBTEACH)**

Item	When learning <school science> topics at school, how often do the following activities occur?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST098Q01TA	Students are given opportunities to explain their ideas.	-0.83337	0.97787	-0.23677	-0.74110	0.67430
ST098Q02TA	Students spend time in the laboratory doing practical experiments.	0.46050	1.06306	-0.49034	-0.57272	0.80028
ST098Q03NA	Students are required to argue about science questions.	0.08387	0.81230	-0.19579	-0.61652	1.17948
ST098Q05TA	Students are asked to draw conclusions from an experiment they have conducted.	-0.10179	0.89409	-0.17404	-0.72005	1.10195
ST098Q06TA	The teacher explains how a <school science> idea can be applied to a number of different phenomena (e.g. the movement of objects, substances with similar properties).	-0.50277	1.01857	-0.16747	-0.85110	0.86825
ST098Q07TA	Students are allowed to design their own experiments.	0.46842	0.46246	-0.15807	-0.30440	1.05809
ST098Q08NA	There is a class debate about investigations.	0.23539	0.67936	-0.16805	-0.51131	1.19736
ST098Q09TA	The teacher clearly explains the relevance of <broad science> concepts to our lives.	-0.36377	0.89348	-0.20540	-0.68808	0.87390

For ST100, students responded on a four-point Likert scale with the categories “every lesson”, “most lessons”, “some lessons” and “never or hardly ever”. As a result, the responses had to be reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of teacher support in science classes. Table 16.33 shows the item wording, international item parameters and item fit for TEACHSUP.

**Table 16.33 Item parameters for Teacher support in a science classes (TEACHSUP)**

Item	How often do these things happen in your <school science> lessons?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST100Q01TA	The teacher shows an interest in every student's learning.	-0.00750	1.26155	-0.08523	-1.17632	0.77330
ST100Q02TA	The teacher gives extra help when students need it.	-0.03532	1.25157	-0.04839	-1.20318	1.09980
ST100Q03TA	The teacher helps students with their learning.	-0.01039	1.10086	-0.02113	-1.07973	1.32146
ST100Q04TA	The teacher continues teaching until the students understand.	0.04437	1.13059	-0.07816	-1.05242	1.01506
ST100Q05TA	The teacher gives students an opportunity to express opinions.	0.01687	1.22992	-0.10423	-1.12570	0.79038

For ST103, students responded on a four-point Likert scale with the categories “never or almost never”, “some lessons”, “many lessons”, and “every lesson or almost every lesson”. Table 16.34 shows the item wording, international item parameters and item fit for TDTEACH.

**Table 16.34 Item parameters for Teacher-directed science instruction (TDTEACH)**

Item	How often do these things happen in your lessons for this <school science> course?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST103Q01NA	The teacher explains scientific ideas.	-0.12171	1.31470	-0.29502	-1.01968	0.82588
ST103Q03NA	A whole class discussion takes place with the teacher.	0.27343	1.26280	-0.21721	-1.04559	0.79269
ST103Q08NA	The teacher discusses our questions.	-0.02685	1.09781	-0.07651	-1.02130	1.32030
ST103Q11NA	The teacher demonstrates an idea.	-0.07612	1.16753	-0.12307	-1.04446	1.06113



For ST104, students responded on a four-point Likert scale with the categories “never or almost never”, “some lessons”, “many lessons”, and “every lesson or almost every lesson”. Table 16.35 shows the item wording, international item parameters and item fit for PERFEED.

**Table 16.35 Item parameters for Perceived Feedback (PERFEED)**

Item	How often do these things happen in your lessons for this <school science> course?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST104Q01NA	The teacher tells me how I am performing in this course.	0.12621	2.18594	-0.46816	-1.71778	0.58887
ST104Q02NA	The teacher gives me feedback on my strengths in this <school science> subject.	0.29760	1.68041	-0.15019	-1.53023	0.89077
ST104Q03NA	The teacher tells me in which areas I can still improve.	0.02181	1.64021	-0.11011	-1.53010	1.23510
ST104Q04NA	The teacher tells me how I can improve my performance.	-0.16677	1.66298	-0.14453	-1.51845	1.28301
ST104Q05NA	The teacher advises me on how to reach my learning goals.	-0.15203	1.56291	-0.15248	-1.41044	1.00225

For ST107, students responded on a four-point Likert scale with the categories “never or almost never”, “some lessons”, “many lessons”, and “every lesson or almost every lesson”. Table 16.36 shows the item wording, international item parameters and item fit for ADINST.

**Table 16.36 Item parameters for Adaption of instruction (ADINST)**

Item	How often do these things happen in your lessons for this <school science> course?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST107Q01NA	The teacher adapts the lesson to my class's needs and knowledge.	-0.00130	1.32590	-0.14690	-1.17900	0.99511
ST107Q02NA	The teacher provides individual help when a student has difficulties understanding a topic or task.	-0.15312	1.33032	-0.14904	-1.18128	1.05697
ST107Q03NA	The teacher changes the structure of the lesson on a topic that most students find difficult to understand.	0.17210	1.21377	-0.08922	-1.12455	0.94792

For ST113, students responded on a four-point Likert scale with the categories “strongly agree”, “agree”, “disagree”, and “strongly disagree”. Therefore, the responses had to be reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of instrumental motivation. INSTSCIE was used in PISA 2006 (ID in 2006: ST35) and thus allows for a trend comparison between PISA 2006 and PISA 2015. Table 16.37 shows the item wording, international item parameters and item fit for INSTSCIE.

**Table 16.37 Item parameters for Instrumental motivation (INSTSCIE)**

Item	How much do you agree with the statements below?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST113Q01TA	Making an effort in my <school science> subject(s) is worth it because this will help me in the work I want to do later on.	-0.12727	1.84275	0.31828	-2.16103	0.94547
ST113Q02TA	What I learn in my <school science> subject(s) is important for me because I need this for what I want to do later on.	0.11242	1.91144	0.17816	-2.08960	1.28323
ST113Q03TA	Studying my <school science> subject(s) is worthwhile for me because what I learn will improve my career prospects.	0.01054	1.95128	0.23715	-2.18843	1.13179
ST113Q04TA	Many things I learn in my <school science> subject(s) will help me to get a job.	0.17985	2.01656	0.21798	-2.23454	0.86955

### **Students' motivation**

New questions were developed for PISA 2015 addressing test anxiety (ANXTEST, ST118) and achievement motivation (MOTIVAT, ST119). Students gave statements about themselves on a four-point Likert scale with the answering categories “strongly agree”, “agree”, “disagree”, and “strongly disagree”. Tables 16.38 and 16.39 contain the scales' reliabilities (Cronbach's Alpha) across all participating OECD and partner countries and economies, respectively.

**Table 16.38 Scale reliabilities for ANXTEST and MOTIVAT in OECD countries**

	ANXTEST	MOTIVAT
Australia	0.852	0.845
Austria	0.829	0.790
Belgium	0.835	0.786
Canada	0.856	0.846
Chile	0.796	0.807
Czech Republic	0.822	0.768
Denmark	0.829	0.841
Estonia	0.830	0.797
Finland	0.808	0.834
France	0.831	0.783
Germany	0.802	0.795
Greece	0.750	0.734
Hungary	0.820	0.788
Iceland	0.895	0.838
Ireland	0.820	0.816
Israel	0.802	0.828
Italy	0.813	0.758
Japan	0.803	0.836
Korea	0.856	0.852
Latvia	0.812	0.797
Luxembourg	0.835	0.820
Mexico	0.803	0.717
Netherlands	0.833	0.753
New Zealand	0.846	0.864
Norway	0.872	0.843
Poland	0.839	0.768
Portugal	0.817	0.779
Slovak Republic	0.822	0.798
Slovenia	0.816	0.795
Spain	0.730	0.773
Sweden	0.856	0.830
Switzerland	0.826	0.780
Turkey	0.825	0.840
United Kingdom	0.849	0.834
United States	0.837	0.855

**Table 16.39 Scale reliabilities for ANXTEST and MOTIVAT in partner countries and economies**

	ANXTEST	MOTIVAT
Albania	N/A	N/A
Algeria	N/A	N/A
Argentina	N/A	N/A
B-S-J-G (China)	0.824	0.780
Brazil	0.716	0.667
Bulgaria	0.841	0.825
Colombia	0.617	0.662
Costa Rica	0.711	0.698
Croatia	0.813	0.773
Cyprus*	0.799	0.798
Dominican Republic	0.705	0.717
FYROM	N/A	N/A
Georgia	N/A	N/A
Hong Kong (China)	0.872	0.831
Indonesia	N/A	N/A
Jordan	N/A	N/A
Kazakhstan	N/A	N/A
Kosovo	N/A	N/A
Lebanon	N/A	N/A
Lithuania	0.830	0.827
Macao (China)	0.845	0.770
Malaysia	0.730	0.845
Malta	N/A	N/A
Moldova	N/A	N/A
Montenegro	0.846	0.804
Peru	0.654	0.695
Qatar	0.780	0.872
Romania	N/A	N/A
Russia	0.814	0.814
Singapore	0.827	0.827
Chinese Taipei	0.839	0.812
Thailand	0.837	0.753
Trinidad and Tobago	N/A	N/A
Tunisia	0.713	0.782
United Arab Emirates	0.762	0.850
Uruguay	0.741	0.729
Viet Nam	N/A	N/A

\* See note under Table 16.6.

Note: N/A indicates that the question has not been administered in the country.



Tables 16.40 and 16.41 show the item wording, international item parameters and item fit for ANXTEST and MOTIVAT, respectively.

**Table 16.40 Item parameters for Test Anxiety (ANXTEST)**

Item	To what extent do you disagree or agree with the following statements about yourself?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST118Q01NA	I often worry that it will be difficult for me taking a test.	-0.05038	1.16536	0.18421	-1.34957	1.16699
ST118Q02NA	I worry that I will get poor <grades> at school.	-0.30152	1.01826	0.22357	-1.24184	1.00140
ST118Q03NA	Even if I am well prepared for a test I feel very anxious.	-0.01720	1.00922	0.13716	-1.14639	1.15496
ST118Q04NA	I get very tense when I study for a test.	0.36492	1.19985	-0.05589	-1.14396	0.96393
ST118Q05NA	I get nervous when I don't know how to solve a task at school.	0.04046	1.16225	0.08846	-1.25071	0.71272

**Table 16.41 Item parameters for Achievement motivation (MOTIVAT)**

Item	To what extent do you disagree or agree with the following statements about yourself?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST119Q01NA	I want top <grades> in most or all of my courses.	-0.15045	1.04968	0.19424	-1.24392	1.25562
ST119Q02NA	I want to be able to select from among the best opportunities available when I graduate.	-0.59253	0.73268	0.61224	-1.34492	1.03250
ST119Q03NA	I want to be the best, whatever I do.	0.18966	1.25665	-0.03152	-1.22513	1.07198
ST119Q04NA	I see myself as an ambitious person.	0.14552	1.75488	0.37566	-2.13054	0.43402
ST119Q05NA	I want to be one of the best students in my class.	0.44301	1.18145	0.03728	-1.21872	1.20588

### **Parental support**

Students were asked about their perceived emotional support from their parents using a newly developed question (ST123) that used a four-point Likert scale with the answering categories “strongly agree”, “agree”, “disagree”, and “strongly disagree”. It included items on whether parents are interested in school activities, support the students’ educational efforts and achievements, support students when they are facing difficulties at school and encourage them to be confident. The derived variable EMOSUPS was scaled using the IRT scaling model described above.

Tables 16.42 and 16.43 contain the scales’ reliabilities (Cronbach’s Alpha) across all participating OECD and partner countries and economies, respectively.

**Table 16.42 Scale reliabilities for the Parental support index in OECD countries**

	EMOSUPS
Australia	0.868
Austria	0.794
Belgium	0.831
Canada	0.872
Chile	0.912
Czech Republic	0.801
Denmark	0.877
Estonia	0.850
Finland	0.894
France	0.840
Germany	0.820
Greece	0.784
Hungary	0.813
Iceland	0.911
Ireland	0.880
Israel	N/A
Italy	0.789
Japan	0.855
Korea	0.889
Latvia	0.861
Luxembourg	0.850
Mexico	0.925
Netherlands	0.847
New Zealand	0.894
Norway	0.888
Poland	0.836
Portugal	0.856
Slovak Republic	0.853
Slovenia	0.761
Spain	0.847
Sweden	0.880
Switzerland	0.825
Turkey	0.856
United Kingdom	0.884
United States	0.871

Note: N/A indicates that the question has not been administered in the country.

**Table 16.43 Scale reliabilities for the Parental support index in partner countries and economies**

	EMOSUPS
Albania	N/A
Algeria	N/A
Argentina	N/A
B-S-J-G (China)	0.788
Brazil	0.818
Bulgaria	0.844
Colombia	0.863
Costa Rica	0.889
Croatia	0.797
Cyprus*	0.830
Dominican Republic	0.882
FYROM	N/A
Georgia	N/A
Hong Kong (China)	0.804
Indonesia	N/A
Jordan	N/A
Kazakhstan	N/A
Kosovo	N/A
Lebanon	N/A
Lithuania	0.850
Macao (China)	0.813
Malaysia	0.731
Malta	N/A
Moldova	N/A
Montenegro	0.762
Peru	0.822
Qatar	0.867
Romania	N/A
Russia	0.806
Singapore	0.851
Chinese Taipei	0.851
Thailand	0.771
Trinidad and Tobago	N/A
Tunisia	0.731
United Arab Emirates	0.816
Uruguay	0.867
Viet Nam	N/A

\* See note under Table 16.6.

Note: N/A indicates that the question has not been administered in the country.

Table 16.44 shows the item wording, international item parameters and item fit for EMOSUPS.

**Table 16.44 Item parameters for Parents emotional support (EMOSUPS)**

Item	Thinking about the <this academic year>; to what extent do you agree or disagree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST123Q01NA	My parents are interested in my school activities.	-0.06068	0.94571	0.96986	-1.91557	0.74465
ST123Q02NA	My parents support my educational efforts and achievements.	-0.14486	0.96658	0.69472	-1.66130	1.15171
ST123Q03NA	My parents support me when I am facing difficulties at school.	0.13633	1.13270	0.53417	-1.66687	1.14779
ST123Q04NA	My parents encourage me to be confident.	0.05811	0.99182	0.60103	-1.59285	0.95585

### Science-related dispositions

Three questions were included to measure science-related dispositions: Science self-efficacy (ST129), epistemological beliefs about science (ST131), and students' science activities (ST146). Tables 16.45 and 16.46 contain the scales' reliabilities (Cronbach's Alpha) across all participating OECD and partner countries and economies, respectively.



Table 16.45 Scale reliabilities for indices on Science related dispositions in OECD countries

	SCIEEFF	EPIST	SCIEACT
Australia	0.907	0.900	0.912
Austria	0.880	0.877	0.908
Belgium	0.880	0.863	0.912
Canada	0.898	0.907	0.924
Chile	0.884	0.900	0.915
Czech Republic	0.859	0.860	0.918
Denmark	0.879	0.912	0.891
Estonia	0.865	0.864	0.904
Finland	0.889	0.903	0.921
France	0.887	0.863	0.909
Germany	0.879	0.850	0.911
Greece	0.865	0.805	0.928
Hungary	0.879	0.834	0.937
Iceland	0.936	0.938	0.915
Ireland	0.873	0.817	0.886
Israel	0.889	0.891	0.946
Italy	0.859	0.840	0.911
Japan	0.913	0.902	0.906
Korea	0.933	0.932	0.931
Latvia	0.828	0.859	0.907
Luxembourg	0.891	0.867	0.923
Mexico	0.885	0.874	0.912
Netherlands	0.895	0.864	0.905
New Zealand	0.901	0.879	0.910
Norway	0.921	0.910	0.926
Poland	0.861	0.883	0.890
Portugal	0.909	0.899	0.925
Slovak Republic	0.892	0.882	0.937
Slovenia	0.863	0.869	0.914
Spain	0.886	0.880	0.911
Sweden	0.915	0.918	0.927
Switzerland	0.880	0.860	0.909
Turkey	0.892	0.919	0.941
United Kingdom	0.902	0.896	0.902
United States	0.900	0.919	0.927

Table 16.46 Scale reliabilities for indices on Science related dispositions in partner countries and economies

	SCIEEFF	EPIST	SCIEACT
Albania	0.822	0.695	N/A
Algeria	0.734	0.707	N/A
Argentina	0.838	0.854	N/A
B-S-J-G (China)	0.891	0.857	0.922
Brazil	0.904	0.873	0.938
Bulgaria	0.888	0.887	0.925
Chinese Taipei	0.917	0.934	0.915
Colombia	0.877	0.858	0.912
Costa Rica	0.888	0.895	0.920
Croatia	0.884	0.876	0.922
Cyprus*	0.904	0.875	0.941
Dominican Republic	0.895	0.913	0.936
FYROM	0.860	0.806	N/A
Georgia	0.835	0.823	N/A
Hong Kong (China)	0.915	0.921	0.937
Indonesia	0.835	0.683	N/A
Jordan	0.840	0.853	N/A
Kazakhstan	0.858	0.829	N/A
Kosovo	0.840	0.790	N/A
Lebanon	0.755	0.731	N/A
Lithuania	0.875	0.906	0.922
Macao (China)	0.887	0.850	0.902
Malaysia	0.888	0.833	0.918
Malta	0.873	0.828	N/A
Moldova	0.815	0.751	N/A
Montenegro	0.906	0.897	0.931
Peru	0.854	0.884	0.909
Qatar	0.898	0.897	0.934
Romania	0.789	0.713	N/A
Russia	0.899	0.882	0.928
Singapore	0.883	0.883	0.917
Thailand	0.885	0.866	0.913
Trinidad and Tobago	0.841	0.832	N/A
Tunisia	0.846	0.798	0.879
United Arab Emirates	0.886	0.874	0.925
Uruguay	0.889	0.911	0.926
Viet Nam	0.782	0.685	N/A

\* See note under Table 16.6.

Note: N/A indicates that the question has not been administered in the country.

Science self-efficacy (ST129) is a trend question that was taken from PISA 2006 (ID in 2006: ST17). Students were asked to rate how they would perform in different science tasks, using a four-point answering scale with the categories "I could do this easily", "I could do this with a bit of effort", "I would struggle to do this on my own", and "I couldn't do this". As a result, the responses had to be reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of science self-efficacy. The derived variable SCIEEFF was scaled using the IRT scaling model described above, thus allowing for a trend comparison between PISA 2006 and PISA 2015. Table 16.47 shows the item wording, international item parameters and item fit for SCIEEFF.

Table 16.47 Item parameters for Science self-efficacy (SCIEEFF)

Item	How easy do you think it would be for you to perform the following tasks on your own?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST129Q01TA	Recognise the science question that underlies a newspaper report on a health issue.	-0.16940	0.98685	0.30908	-1.29594	0.93845
ST129Q02TA	Explain why earthquakes occur more frequently in some areas than in others.	-0.27092	0.83348	0.16974	-1.00323	0.92431
ST129Q03TA	Describe the role of antibiotics in the treatment of disease.	0.06516	0.88992	0.10362	-0.99354	1.00384
ST129Q04TA	Identify the science question associated with the disposal of garbage.	0.00601	0.93480	0.13846	-1.07326	1.04883
ST129Q05TA	Predict how changes to an environment will affect the survival of certain species.	-0.03415	0.82526	0.13232	-0.95758	1.13443
ST129Q06TA	Interpret the scientific information provided on the labelling of food items.	-0.04337	0.91786	0.12501	-1.04287	0.98109
ST129Q07TA	Discuss how new evidence can lead you to change your understanding about the possibility of life on Mars.	0.28023	0.80702	0.13201	-0.93903	0.97553
ST129Q08TA	Identify the better of two explanations for the formation of acid rain.	0.14654	0.78166	0.13256	-0.91422	0.99352

Epistemological beliefs about science were measured with a new question about students' views on scientific approaches (ST131). Students answered on a four-point Likert scale with the answering categories "strongly agree", "agree", "disagree", and "strongly disagree". The derived variable EPIST was scaled using the IRT scaling model described above. Table 16.48 shows the item wording, international item parameters and item fit for EPIST.

Table 16.48 Item parameters for Epistemological beliefs (EPIST)

Item	How much do you disagree or agree with the statements below?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST131Q01NA	A good way to know if something is true is to do an experiment.	0.00900	0.69269	1.00678	-1.69947	0.83989
ST131Q03NA	Ideas in <broad science> sometimes change.	0.12064	1.37107	0.58817	-1.95924	1.11811
ST131Q04NA	Good answers are based on evidence from many different experiments.	-0.11558	1.01482	0.58431	-1.59913	1.16975
ST131Q06NA	It is good to try experiments more than once to make sure of your findings.	-0.19914	0.95392	0.54680	-1.50072	1.06412
ST131Q08NA	Sometimes <broad science> scientists change their minds about what is true in science.	0.11261	1.37343	0.58717	-1.96059	0.96138
ST131Q11NA	The ideas in <broad science> science books sometimes change.	0.11386	1.39472	0.60798	-2.00270	0.84676

Another trend question from PISA 2006 (ID in 2006: ST19) addressed students' science activities (ST146). Students were asked how often they engaged in science-related activities on a four-point scale with the answering categories "very often", "regularly", "sometimes", and "never or hardly ever". Therefore, the responses had to be reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of students' science activities. The derived variable SCIEACT was scaled using the IRT scaling model described above, thus allowing for a trend comparison between PISA 2006 and PISA 2015. Table 16.49 shows the item wording, international item parameters and item fit for SCIEACT.

Table 16.49 Item parameters for Science activities (SCIEACT)

Item	How often do you do these things?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
ST146Q01TA	Watch TV programmes about <broad science>	-0.75036	1.87968	-0.83248	-1.04720	0.64488
ST146Q02TA	Borrow or buy books on <broad science> topics	-0.01962	1.10537	-0.35079	-0.75459	1.10371
ST146Q03TA	Visit web sites about <broad science> topics	-0.17128	1.23499	-0.44357	-0.79142	0.72994
ST146Q04TA	Read <broad science> magazines or science articles in newspapers	-0.37920	1.26074	-0.38471	-0.87602	0.85784
ST146Q05TA	Attend a <science club>	0.45931	0.42171	-0.06008	-0.36164	0.83529
ST146Q06NA	Simulate natural phenomena in computer programs/virtual labs	0.16648	0.82516	-0.11161	-0.71355	1.50118
ST146Q07NA	Simulate technical processes in computer programs/virtual labs	0.15594	0.78812	-0.12517	-0.66295	1.43343
ST146Q08NA	Visit web sites of ecology organisations	0.07336	0.98205	-0.22220	-0.75984	1.14309
ST146Q09NA	Follow news of science, environmental, or ecology organizations via blogs and microblogging	-0.05048	0.96640	-0.24713	-0.71927	0.86875



## SCHOOL QUESTIONNAIRE DERIVED VARIABLES

The PISA 2015 School Questionnaire consisted mainly of trend questions used in previous cycles. As the major domain of the 2015 cycles was once again science, some scales focused on science-specific aspects of learning context on a school level. However, no trend scales were reported in both 2006 and 2015 cycles. All derived variables are shown in Table 16.50 and described below. Simple questionnaire indices are preceded by those that are based on IRT scaling.

**Table 16.50 Derived variables in the PISA 2015 School Questionnaire**

DV Name	Description	Question no.	Trend to PISA 2006	IRT scaling
SCHSIZE	School Size	SC002		
CLSIZE	Class Size	SC003		
RATCMP1	Index of computer availability	SC004		
RATCMP2	Index of computers connected to the Internet	SC004		
LEAD	Educational leadership	SC009		YES
LEADCOM	Curricular development	SC009		YES
LEADINST	Instructional leadership	SC009		YES
LEADPD	Professional development	SC009		YES
LEADTCH	Teachers participation	SC009		YES
RESPCUR	Responsibility for curriculum	SC010		
RESPRES	Responsibility for resources	SC010		
SCHAUT	School autonomy	SC010		
TEACHPART	Teacher participation	SC010		
SCHLTYPE	School Ownership	SC013, SC016		
EDUSHORT	Shortage of educational material	SC017		YES
STAFFSHORT	Shortage of educational staff	SC017		YES
PROAT5AB	Proportion of all teachers ISCED LEVEL 5A Bachelor	SC018		
PROAT5AM	Proportion of all teachers ISCED LEVEL 5A Master	SC018		
PROAT6	Proportion of all teachers ISCED LEVEL 6	SC018		
PROATCE	Proportion of all teachers fully certified	SC018		
TOTAT	Total number of all teachers at school	SC018		
STRATIO	Student teacher ratio	SC018, SC002		
PROSTAT	Proportion of science teachers by all teachers	SC018, SC019		
PROSTCE	Proportion of science teachers fully certified	SC019		
PROSTMAS	Proportion of science teachers with ISCED level 5A and a major in science	SC019		
TOTST	Total number of science teachers at school	SC019		
CREACTIV	Creative extra-curricular activities	SC053		
SCIRES	Science specific resources	SC059		
STUBEHA	Student-related factors affecting school climate	SC061		YES
TEACHBEHA	Teacher-related factors affecting school climate	SC061		YES

## Simple questionnaire indices

### School size

The index of school size (SCHSIZE) contains the total enrolment at school. It is based on the enrolment data provided by the school principal, summing the number of girls and boys at a school (SC002). This index was calculated in 2015 and in all previous cycles.

### Class size

The average class size (CLSIZE) is derived from one of nine possible categories in question SC003, ranging from “15 students or fewer” to “More than 50 students”.

### Availability of computers

School principals were asked to report the number of computers available at school (SC004). The index of availability of computers (RATCMP1) is the ratio of computers available to 15-year olds for educational purposes to the total number of students in the modal grade for 15-year olds. The index RATCMP2 was calculated as the ratio of number of computers available to 15-year olds for educational purposes to the number of these computers that were connected to the internet.

A new index was built in 2015 to reflect the schools’ science-specific resources (SCIRES). It was constructed by summing up the principals’ answers to SC059 (yes/no question).



### **School responsibility**

As in previous cycles, school responsibility for curriculum and resources as well as school autonomy and teacher participation was addressed in question SC010. An index of the relative level of responsibility of school staff in allocating resources (RESPRES) was derived from six items of the school principals' report regarding who had considerable responsibility for tasks related to resource allocation ("selecting teachers for hire", "firing teachers", "establishing teachers' starting salaries", "determining teachers' salary increases", "formulating the school budget", "deciding on budget allocations within the school"). The index was calculated on the basis of the ratio of "yes" responses for school governing board, principal or teachers to "yes" responses for regional/local education authority or national educational authority. Higher values on the scale indicated relatively higher levels of school responsibility in this area. The index was standardised to having an OECD mean of '0' and a standard deviation of '1' for the pooled data set with equally weighted country samples. This index was also created in the 2006, 2009 and 2012 PISA cycles.

An index of the relative level of responsibility of school staff in issues relating to curriculum and assessment (RESPCUR) was computed from the school principal's report regarding who had responsibility for four aspects of curriculum and assessment, namely "establishing student assessment policies", "choosing which textbooks are used", "determining course content", and "deciding which courses are offered". The index was calculated on the basis of the ratio of "yes" responses for school governing board, principal or teachers on the one hand to "yes" responses for regional/local education authority or national educational authority on the other hand. Higher values indicated relatively higher levels of school responsibility in this area. The index was standardised to having an OECD mean of '0' and a standard deviation of '1' for the pooled data with equally weighted country samples). This index was also created in all previous PISA cycles, although in PISA 2009 the variable name was RESPCURR.

### **School type**

Schools are classified as either public or private according to whether a private entity or a public agency has the ultimate power for decision making concerning its affairs. As in previous PISA surveys, the index on school type (SCHLTYPE) has three categories, based on two questions: SC013 asks if the school is a public or a private school, SC016 asks about the source of resources. This index was calculated in 2015 and in all previous cycles. In 2009 the variable name was SCHTYPE.

### **Quantity of teaching staff at school**

Principals were asked to report the total number of teachers at their school (TOTAT) and provide additional information on how many of the staff was full-time and part-time employed teachers qualified at different ISCED levels (SC018).

The proportion of fully certified teachers (PROATCE) was computed by dividing the number of fully certified teachers by the total number of teachers.

The proportion of teachers with an ISCED 5A bachelor qualification (PROAT5AB) was calculated by dividing the number of these teachers by the total number of teachers.

The proportion of teachers with an ISCED 5A master qualification (PROAT5AM) was calculated by dividing the number of these teachers by the total number of teachers.

The proportion of teachers with an ISCED level 6 qualification (PROAT6) was calculated by dividing the number of these teachers by the total number of teachers.

The student-teacher ratio (STRATIO) was obtained by dividing the number of enrolled students (SC002) by the total number of teachers (TOTAT).

An additional question (SC019) asked about the number of science teachers at the school, including information about full-time or part-time employment and the respective ISCED level qualification of these science teachers.

The proportion of science teachers (PROSTAT) was computed by dividing the number of science teachers by the total number of teachers.

The proportion of fully certified science teachers (PROSTCE) was computed by dividing the number of fully certified science teachers by the total number of teachers.



The proportion of science teachers with an ISCED 5A qualification and a major in science (PROSTMAS) was calculated by dividing the number of these teachers by the total number of science teachers.

### **Extra-curricular activities at school**

School principals were asked to report what extra-curricular activities their schools offered to 15-year old students (SC053). The index of creative extra-curricular activities at school (CREACTIV) was computed as the total number of the following activities that occurred at school: i) band, orchestra or choir; ii) school play or school musical; and iii) art club or art activities.

### **Derived variables based on IRT Scaling**

The School Questionnaire provided data for nine scaled indices which will be presented along with the item content and parameters in the following. Tables 16.51 and 16.52 contain the scale reliabilities (Cronbach's Alpha coefficients) for all participating OECD and partner countries and economies, respectively.

**Table 16.51 Scale reliabilities for School Questionnaire indices in OECD countries**

	LEAD	LEADCOM	LEADINST	LEADPD	LEADTCH	EDUSHORT	STAFFSHORT	STUBEHA	TEACHBEHA
Australia	0.914	0.790	0.795	0.811	0.814	0.869	0.799	0.850	0.805
Austria	0.902	0.761	0.760	0.826	0.794	0.838	0.646	0.804	0.749
Belgium	0.902	0.726	0.761	0.834	0.789	0.829	0.670	0.801	0.782
Canada	0.899	0.766	0.792	0.767	0.760	0.841	0.765	0.836	0.810
Chile	0.912	0.804	0.782	0.815	0.691	0.842	0.815	0.865	0.802
Czech Republic	0.893	0.701	0.754	0.848	0.765	0.782	0.642	0.799	0.696
Denmark	0.869	0.728	0.701	0.822	0.782	0.876	0.755	0.793	0.813
Estonia	0.856	0.697	0.695	0.814	0.650	0.781	0.767	0.692	0.769
Finland	0.900	0.725	0.756	0.834	0.695	0.857	0.680	0.761	0.781
France	0.902	0.749	0.724	0.868	0.769	0.834	0.713	0.766	0.784
Germany	0.888	0.687	0.747	0.789	0.735	0.846	0.701	0.771	0.631
Greece	0.903	0.703	0.790	0.881	0.833	0.878	0.653	0.818	0.768
Hungary	0.888	0.733	0.734	0.837	0.669	0.825	0.534	0.821	0.722
Iceland	0.894	0.754	0.735	0.829	0.720	0.824	0.717	0.763	0.782
Ireland	0.897	0.721	0.757	0.754	0.754	0.870	0.719	0.760	0.842
Israel	0.899	0.762	0.693	0.813	0.796	0.834	0.811	0.670	0.821
Italy	0.886	0.736	0.682	0.810	0.779	0.864	0.689	0.767	0.807
Japan	0.840	0.755	0.656	0.732	0.687	0.903	0.732	0.767	0.674
Korea	0.923	0.714	0.773	0.834	0.869	0.880	0.701	0.832	0.806
Latvia	0.860	0.652	0.709	0.804	0.764	0.815	0.751	0.752	0.758
Luxembourg	0.887	0.749	0.760	0.641	0.863	0.831	0.745	0.773	0.765
Mexico	0.906	0.821	0.746	0.759	0.785	0.906	0.721	0.791	0.845
Netherlands	0.888	0.716	0.705	0.857	0.818	0.789	0.716	0.794	0.706
New Zealand	0.894	0.669	0.709	0.798	0.776	0.816	0.741	0.822	0.814
Norway	0.903	0.797	0.760	0.799	0.758	0.837	0.695	0.768	0.761
Poland	0.860	0.665	0.678	0.811	0.721	0.835	0.687	0.753	0.812
Portugal	0.905	0.740	0.826	0.805	0.795	0.868	0.710	0.803	0.819
Slovak Republic	0.893	0.644	0.699	0.848	0.775	0.808	0.608	0.777	0.722
Slovenia	0.912	0.761	0.843	0.819	0.717	0.806	0.765	0.748	0.718
Spain	0.863	0.657	0.726	0.789	0.737	0.901	0.726	0.787	0.832
Sweden	0.900	0.741	0.747	0.823	0.662	0.807	0.824	0.736	0.791
Switzerland	0.861	0.698	0.694	0.823	0.763	0.810	0.647	0.797	0.739
Turkey	0.909	0.679	0.818	0.755	0.867	0.905	0.804	0.802	0.751
United Kingdom	0.897	0.780	0.751	0.829	0.792	0.833	0.714	0.801	0.806
United States	0.916	0.737	0.730	0.780	0.795	0.854	0.840	0.797	0.869

**Table 16.52 Scale reliabilities for School Questionnaire in partner countries and economies**

	LEAD	LEADCOM	LEADINST	LEADPD	LEADTCH	EDUSHORT	STAFFSHORT	STUBEHA	TEACHBEHA
Albania	0.844	0.702	0.612	0.734	0.733	0.859	0.736	0.779	0.739
Algeria	0.918	0.673	0.831	0.856	0.823	0.819	0.682	0.787	0.664
Argentina	0.893	0.777	0.722	0.796	0.704	0.842	0.746	0.758	0.809
B-S-J-G (China)	0.888	0.680	0.731	0.755	0.807	0.939	0.885	0.959	0.906
Brazil	0.909	0.780	0.757	0.806	0.789	0.848	0.760	0.833	0.847
Bulgaria	0.902	0.722	0.786	0.788	0.823	0.762	0.693	0.875	0.879
Colombia	0.929	0.835	0.795	0.856	0.765	0.892	0.824	0.860	0.839
Costa Rica	0.916	0.749	0.735	0.866	0.808	0.869	0.813	0.858	0.826
Croatia	0.918	0.716	0.812	0.871	0.763	0.813	0.642	0.825	0.820
Cyprus*	0.882	0.674	0.707	0.854	0.806	0.894	0.868	0.768	0.680
Dominican Republic	0.867	0.745	0.727	0.662	0.655	0.807	0.753	0.763	0.761
FYROM	0.901	0.763	0.812	0.754	0.804	0.854	0.756	0.794	0.769
Georgia	0.861	0.627	0.621	0.727	0.769	0.860	0.741	0.865	0.848
Hong Kong (China)	0.914	0.780	0.759	0.834	0.832	0.885	0.821	0.720	0.820
Indonesia	0.908	0.774	0.747	0.772	0.820	0.885	0.792	0.667	0.578
Jordan	0.869	0.618	0.702	0.741	0.782	0.905	0.854	0.833	0.819
Kazakhstan	0.845	0.627	0.577	0.720	0.750	0.874	0.823	0.913	0.939
Kosovo	0.886	0.715	0.679	0.783	0.783	0.789	0.756	0.844	0.793
Lebanon	0.855	0.745	0.719	0.692	0.721	0.890	0.739	0.811	0.828
Lithuania	0.892	0.684	0.687	0.843	0.782	0.803	0.613	0.776	0.788
Macao (China)	0.868	0.716	0.611	0.818	0.773	0.911	0.901	0.945	0.924
Malaysia	0.944	0.815	0.844	0.851	0.873	0.876	0.827	0.860	0.820
Malta	0.784	0.614	0.598	0.699	0.653	0.815	0.739	0.794	0.770
Moldova	0.820	0.540	0.642	0.675	0.782	0.767	0.735	0.821	0.837
Montenegro	0.902	0.743	0.759	0.786	0.793	0.889	0.654	0.757	0.806
Peru	0.930	0.816	0.797	0.796	0.818	0.882	0.769	0.829	0.873
Qatar	0.880	0.713	0.663	0.764	0.813	0.877	0.856	0.762	0.798
Romania	0.854	0.626	0.736	0.579	0.681	0.796	0.703	0.807	0.794
Russia	0.889	0.762	0.714	0.809	0.781	0.874	0.799	0.851	0.889
Singapore	0.917	0.802	0.766	0.844	0.799	0.813	0.854	0.778	0.761
Chinese Taipei	0.928	0.811	0.782	0.823	0.881	0.866	0.713	0.929	0.858
Thailand	0.932	0.807	0.817	0.841	0.879	0.884	0.767	0.803	0.798
Trinidad and Tobago	0.876	0.656	0.739	0.772	0.726	0.842	0.820	0.829	0.839
Tunisia	0.842	0.526	0.628	0.735	0.797	0.827	0.733	0.840	0.821
United Arab Emirates	0.889	0.745	0.681	0.773	0.777	0.930	0.894	0.849	0.856
Uruguay	0.884	0.697	0.681	0.795	0.775	0.865	0.814	0.825	0.819
Viet Nam	0.897	0.642	0.755	0.823	0.738	0.846	0.711	0.699	0.737

\* See note under Table 16.6.

### **School leadership**

A question on school leadership was developed for PISA 2012 and partially taken up again for PISA 2015. Question SC009 with 13 items asks about school leadership. The results provided data for five scaled indices. Principals were asked to indicate the frequency of the listed activities and behaviours in their school during the last academic year. The six response categories were “did not occur”, “1-2 times during the year”, “3-4 times during the year”, “once a month”, “once a week”, to “more than once a week”. The overall scale for leadership (LEAD) consists of all 13 items. Table 16.53 shows the item wording, international item parameters and item fit for LEAD.

**Table 16.53 Item parameters for Educational leadership (LEAD)**

Item	Below are statements about your management of this school. Please indicate the frequency of the following activities and behaviours in your school during <the last academic year>.	Parameter estimates						
		beta	d_1	d_2	d_3	d_4	d_5	alpha
SC009Q01TA	I use student performance results to develop the school's educational goals.	0.46464	2.35073	0.32178	-0.69840	-1.35309	-0.62102	0.75818
SC009Q02TA	I make sure that the professional development activities of teachers are in accordance with the teaching goals of the school.	0.29463	2.09230	0.18305	-0.38615	-1.08206	-0.80714	0.83482
SC009Q03TA	I ensure that teachers work according to the school's educational goals.	-0.11346	1.86425	0.25513	-0.34342	-0.95865	-0.81732	1.00750
SC009Q04TA	I promote teaching practices based on recent educational research.	0.32348	1.44205	0.16228	-0.19596	-0.93725	-0.47112	0.87299
SC009Q05TA	I praise teachers whose students are actively participating in learning.	-0.01904	1.38715	0.29741	-0.14263	-0.85219	-0.68974	0.98060
SC009Q06TA	When a teacher has problems in his/her classroom, I take the initiative to discuss matters.	-0.13401	1.13879	0.37206	-0.07387	-0.67910	-0.75787	1.00091
SC009Q07TA	I draw teachers' attention to the importance of pupils' development of critical and social capacities.	0.05311	1.17565	0.36168	-0.10448	-0.67021	-0.76264	1.47738
SC009Q08TA	I pay attention to disruptive behaviour in classrooms.	-0.38714	0.76147	0.40453	-0.06023	-0.48927	-0.61649	0.92058
SC009Q09TA	I provide staff with opportunities to participate in school decision-making.	-0.18983	1.44581	0.53346	-0.12085	-0.93344	-0.92498	0.91883
SC009Q10TA	I engage teachers to help build a school culture of continuous improvement.	-0.17508	1.24219	0.43019	-0.12964	-0.73994	-0.80281	1.37113
SC009Q11TA	I ask teachers to participate in reviewing management practices.	0.39472	1.47123	0.09030	-0.04154	-0.84679	-0.67320	0.79238
SC009Q12TA	When a teacher brings up a classroom problem, we solve the problem together.	-0.32621	1.18322	0.38598	-0.11159	-0.66445	-0.79317	1.07053
SC009Q13TA	I discuss the school's academic goals with teachers at faculty meetings.	0.11599	1.75821	0.62338	-0.11506	-1.15915	-1.10738	0.99417

The index LEADCOM reflects how school's goals and curricular development are framed and communicated. The IRT scaling model uses items SC009Q01TA, SC009Q02TA, SC009Q03TA, and SC009Q13TA. Table 16.54 shows the item wording, international item parameters and item fit for LEADCOM.

**Table 16.54 Item parameters for Curricular development (LEADCOM)**

Item	Below are statements about your management of this school. Please indicate the frequency of the following activities and behaviours in your school during <the last academic year>.	Parameter estimates						
		beta	d_1	d_2	d_3	d_4	d_5	alpha
SC009Q01TA	I use student performance results to develop the school's educational goals.	0.32244	2.38253	0.38339	-0.65558	-1.32651	-0.78382	0.88402
SC009Q02TA	I make sure that the professional development activities of teachers are in accordance with the teaching goals of the school.	0.17496	1.88930	0.30071	-0.31267	-0.93355	-0.94378	1.35180
SC009Q03TA	I ensure that teachers work according to the school's educational goals.	-0.31443	2.00895	0.34210	-0.35579	-1.00368	-0.99158	1.19417
SC009Q13TA	I discuss the school's academic goals with teachers at faculty meetings.	-0.25626	2.71719	0.87071	-0.24749	-1.85411	-1.48629	0.57000

The index reflecting instructional leadership (LEADINST) at a school is built by scaling items SC009Q04TA, SC009Q05TA, and SC009Q07TA. Table 16.55 shows the item wording, international item parameters and item fit for LEADINST.

**Table 16.55 Item parameters for Instructional leadership (LEADINST)**

Item	Below are statements about your management of this school. Please indicate the frequency of the following activities and behaviours in your school during <the last academic year>.	Parameter estimates						
		beta	d_1	d_2	d_3	d_4	d_5	alpha
SC009Q04TA	I promote teaching practices based on recent educational research.	0.26577	1.59619	0.21769	-0.21557	-1.00781	-0.59051	0.88207
SC009Q05TA	I praise teachers whose students are actively participating in learning.	-0.09737	1.49024	0.36747	-0.14803	-0.87302	-0.83666	1.12929
SC009Q07TA	I draw teachers' attention to the importance of pupils' development of critical and social capacities.	-0.12589	1.57384	0.43930	-0.17246	-0.92012	-0.92056	0.98864

The index on how instructional improvements and professional development are promoted by the principal (LEADPD) is scaled by using items SC009Q06TA, SC009Q08TA, and SC009Q12TA. Table 16.56 shows the item wording, international item parameters and item fit for LEADPD.

**Table 16.56 Item parameters for Professional development (LEADPD)**

Item	Below are statements about your management of this school. Please indicate the frequency of the following activities and behaviours in your school during <the last academic year>.	Parameter estimates						
		beta	d_1	d_2	d_3	d_4	d_5	alpha
SC009Q06TA	When a teacher has problems in his/her classroom, I take the initiative to discuss matters.	0.20078	1.63228	0.62162	-0.04782	-0.91567	-1.29040	0.92126
SC009Q08TA	I pay attention to disruptive behaviour in classrooms.	-0.17788	1.15397	0.62822	-0.02301	-0.66721	-1.09198	0.86532
SC009Q12TA	When a teacher brings up a classroom problem, we solve the problem together.	-0.02559	1.57517	0.61935	-0.06903	-0.82037	-1.30512	1.21342

The index of teacher participation in leadership (LEADTCH) is reported using items SC009Q09TA, SC009Q10TA, and SC009Q11TA. Table 16.57 shows the item wording, international item parameters and item fit for LEADTCH.

**Table 16.57 Item parameters for Teachers participation (LEADTCH)**

Item	Below are statements about your management of this school. Please indicate the frequency of the following activities and behaviours in your school during <the last academic year>.	Parameter estimates						
		beta	d_1	d_2	d_3	d_4	d_5	alpha
SC009Q09TA	I provide staff with opportunities to participate in school decision-making.	-0.11497	1.90159	0.73619	-0.13752	-1.10972	-1.39054	1.04028
SC009Q10TA	I engage teachers to help build a school culture of continuous improvement.	-0.17367	1.78244	0.64523	-0.17203	-0.98132	-1.27432	1.39244
SC009Q11TA	I ask teachers to participate in reviewing management practices.	0.63711	2.20995	0.18705	-0.06463	-1.25134	-1.08103	0.56728

### School resources

PISA 2015 included a question with eight items about school resources, measuring the school principals' perceptions of potential factors hindering the provision of instruction at school. The four response categories were "not at all", "very little", "to some extent", to "a lot". A similar question was used in previous cycles, but items were reduced and reworded for 2015 focusing on two derived variables. The index on staff shortage (STAFFSHORT) was derived from four items SC017Q01NA, SC017Q02NA, SC017Q03NA, and SC017Q04NA. The index on shortage of educational material (EDUSHORT) was scaled using four items SC017Q05NA, SC017Q06NA, SC017Q07NA, and SC017Q08NA. The items were not reversed for scaling. Tables 16.58 and 16.59 show the item wording, international item parameters and item fit for STAFFSHORT and EDUSHORT, respectively.

**Table 16.58 Item parameters for Shortage of educational material (EDUSHORT)**

Item	Is your school's capacity to provide instruction hindered by any of the following issues?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
SC017Q05NA	A lack of educational material (e.g. textbooks, IT equipment, library or laboratory material).	0.21882	1.59613	0.43175	-2.02788	0.39524
SC017Q06NA	Inadequate or poor quality educational material (e.g. textbooks, IT equipment, library or laboratory material).	0.43446	1.84628	0.30677	-2.15305	0.40730
SC017Q07NA	A lack of physical infrastructure (e.g. building, grounds, heating/cooling, lighting and acoustic systems).	-0.11732	1.23750	0.14076	-1.37826	1.53249
SC017Q08NA	Inadequate or poor quality physical infrastructure (e.g. building, grounds, heating/cooling, lighting and acoustic systems).	-0.05024	1.32658	0.10092	-1.42751	1.66497

**Table 16.59 Item parameters for Shortage of educational staff (STAFFSHORT)**

Item	Is your school's capacity to provide instruction hindered by any of the following issues?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
SC017Q01NA	A lack of teaching staff.	0.06314	0.75909	0.34257	-1.10165	0.73336
SC017Q02NA	Inadequate or poorly qualified teaching staff.	0.13603	1.05954	0.03870	-1.09824	0.92824
SC017Q03NA	A lack of assisting staff.	-0.26505	0.60896	0.21200	-0.82096	0.95589
SC017Q04NA	Inadequate or poorly qualified assisting staff.	0.05843	0.72331	0.05539	-0.77870	1.38251

### School climate

The School Questionnaire included a trend question on school climate (SC061) that had been used in previous cycles with a larger set of items. It measured the school principals' perceptions of the school climate, in particular his or her perceptions of teacher and student behaviour that might influence the provision of instruction at school. The four response categories were "not at all", "very little", "to some extent" and "a lot". For PISA 2015, the items were rearranged to reflect student-related factors (STUBEHA) and teacher-related factors (TEACHBEHA) affecting school climate. The scaling model



used items SC061Q01TA, SC061Q02TA, SC061Q03TA, SC061Q04TA, and SC061Q05TA to reflect STUBEHA, and SC061Q06TA, SC061Q07TA, SC061Q08TA, SC061Q09TA, and SC061Q10TA to reflect TEACHBEHA. Tables 16.60 and 16.61 show the item wording, international item parameters and item fit for STUBEHA and TEACHBEHA, respectively.

**Table 16.60 Item parameters for Student-related factors affecting school climate (STUBEHA)**

Item	In your school, to what extent is the learning of students hindered by the following phenomena?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
SC061Q01TA	Student truancy	-0.46872	1.48863	-0.12469	-1.36395	1.25759
SC061Q02TA	Students skipping classes	-0.28674	1.50293	-0.09281	-1.41012	1.46127
SC061Q03TA	Students lacking respect for teachers	0.08023	1.88225	-0.35878	-1.52347	0.81146
SC061Q04TA	Student use of alcohol or illegal drugs	0.73855	1.14267	-0.51484	-0.62783	0.78086
SC061Q05TA	Students intimidating or bullying other students	0.53229	2.05337	-0.64487	-1.40851	0.68882

**Table 16.61 Item parameters for Teacher-related factors affecting school climate (TEACHBEHA)**

Item	In your school, to what extent is the learning of students hindered by the following phenomena?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
SC061Q06TA	Teachers not meeting individual students' needs	-0.05338	1.63983	-0.03794	-1.60189	1.06092
SC061Q07TA	Teacher absenteeism	0.00094	1.44904	-0.26889	-1.18014	0.88448
SC061Q08TA	Staff resisting change	-0.22931	1.37983	0.04536	-1.42519	1.09578
SC061Q09TA	Teachers being too strict with students	0.43368	2.16129	-0.20726	-1.95403	0.71598
SC061Q10TA	Teachers not being well prepared for classes	-0.00276	1.44495	-0.31509	-1.12986	1.24283

## EDUCATIONAL CAREER QUESTIONNAIRE

The Educational Career Questionnaire (ECQ) is an international option that countries can choose to implement. It is administered to the PISA students after they have completed the Student Questionnaire. As the content of the ECQ changes in every cycle, no trend scales were built for PISA 2015. The derived variables of the ECQ are simple questionnaire indices only. An overview of all derived variables is shown in Table 16.62, and each index is described in the following sections.

**Table 16.62 Derived variables in the optional PISA 2015 Educational Career Questionnaire**

DV Name	Description	Question no.	Trend to PISA 2006	IRT scaling
HADDINST	Total hours of additional instruction	EC001		
SADDINST	Number of learning domains with additional instruction	EC001		
ADDSCIIN	Number of science disciplines and subjects with additional instruction	EC003		
COMSCSUP	Comparison science school lessons and additional instruction support	EC009		
COMSCSTRCO	Comparison science school lessons and additional instruction structuredness content	EC010		
COMSCSTRLE	Comparison science school lessons and additional instruction structuredness lessons	EC010		
COMSCTSREL	Comparison science school lessons and additional instruction teacher-student relation	EC011		
COMMASUP	Comparison mathematics school lessons and additional instruction support	EC019		
COMMSTRCO	Comparison mathematics school lessons and additional instruction structuredness content	EC020		
COMMSTRLE	Comparison mathematics school lessons and additional instruction structuredness lessons	EC020		
COMMATSREL	Comparison mathematics school lessons and additional instruction teacher-student relation	EC021		
SCCHANGE	Number of school changes	EC031, EC032		
CHANGE	Number of changes in educational biography	EC031-EC033		

## Simple questionnaire indices

### Learning time

Question EC001 asks about the hours per week that the student attended any additional instruction, and the subjects that were covered in this additional instruction. The derived variable HADDINST reflects the sum of all hours of additional instruction. The derived variable SADDINST states the number of individual subjects in which a student attends additional lessons.

To focus on science-specific additional instruction (SC003), the derived variable ADDSCIIN reflects the sum of all science disciplines and subjects in which the student attends additional lessons.

### **Instructional quality**

To assess the instructional quality of additional instruction, PISA 2015 included newly developed questions that asked students to compare the quality in regular school lessons to that in their additional instruction. The questions focused on science (EC009/EC010/EC011) and mathematics (EC019/EC020/EC021). For each aspect, the student was asked whether it was more likely to occur in the regular school lessons, the additional instruction, or if there was no difference between the two.

Aspects included a comparison of teacher support in science lessons (COMSCSUP, EC009) and mathematics lessons (COMMASUP, EC019), derived variables are built based on the mean of all answers.

Questions EC010 (for science) and EC020 (for mathematics) asked about the structuredness of the lessons. The respective indicators report the structuredness of content (e.g. pacing, curriculum coherence, COMSCSTRCO) and the structuredness of the lessons (e.g. classroom management, COMSCSTRLE) for science, as well as for mathematics (COMMSTRCO/COMMSTRLE).

In addition, students were asked to compare teacher-student relationships. The respective indicators are COMSCTSREL for science (EC011) and COMMATSREL for mathematics (EC021).

### **Educational pathways**

The Educational Career Questionnaire also included questions about the students' educational pathways within the school system, asking for information on whether students had ever changed schools when attending ISCED 1 (EC031) or ISCED 2 (EC031), as well as whether they had ever changed a study programme (EC033).

The respective indicators summed up the number of school changes in EC031 and EC032 (SCCHANGE) and all three questions reported on the number of overall changes in the educational biography (CHANGE).

## **ICT FAMILIARITY QUESTIONNAIRE**

The ICT Familiarity Questionnaire (ICQ) is an international option that countries can choose to implement. It is administered to the PISA students after they have completed the Student Questionnaire. For PISA 2015, nine derived variables were built, eight of which were scaled using the IRT model described above. Most of the scales were already reported in PISA 2012 but some now include updated items and further theoretical constructs.

An overview of all derived variables is shown in Table 16.63, and each is described in the following sections. Simple questionnaire indices are preceded by those that are based on IRT scaling.

**Table 16.63 Derived variables in the optional PISA 2015 ICT Familiarity Questionnaire**

DV Name	Description	Question no.	Trend to PISA 2006	IRT scaling
ICTHOME	ICT available at Home Index	IC001		
ENTUSE	ICT use outside of school leisure	IC008		YES
ICTSCH	ICT available at School Index	IC009		
HOMESCH	ICT use outside of school for schoolwork	IC010		YES
USESCH	Use of ICT at school in general	IC011		YES
INTICT	Students' ICT Interest	IC013		YES
COMPICT	Students' Perceived ICT Competence	IC014		YES
AUTICT	Students' Perceived Autonomy related to ICT Use	IC015		YES
SOIAICT	Students' ICT as a topic in Social Interaction	IC016		YES

### **Simple questionnaire indices**

#### **Availability and usage of ICT**

The ICQ asked about the availability of ICT at home and if students used it for various purposes. ICTHOME is an index based on the sum of the availability of all items included in IC001.

IC009 asked about the availability of ICT at school, the respective derived variable ICTSCH is calculated as the sum of all items.



## Derived variables based on IRT Scaling

The ICT Familiarity Questionnaire provided data for seven scaled indices which will be presented along with the item content and parameters in the following sections. Tables 16.64 and 16.65 contain the scale reliabilities (Cronbach's Alpha coefficients) for all participating OECD and partner countries and economies, respectively.

Table 16.64 Scale reliabilities for ICT Familiarity Questionnaire indices in OECD countries

	ENTUSE	HOMESCH	USESCH	INTICT	COMPCT	AUTICT	SOIAICT
Australia	0.804	0.906	0.836	0.785	0.848	0.871	0.850
Austria	0.784	0.885	0.857	0.765	0.840	0.840	0.864
Belgium	0.797	0.919	0.910	0.794	0.846	0.811	0.855
Chile	0.831	0.911	0.867	0.797	0.839	0.850	0.859
Czech Republic	0.810	0.901	0.887	0.775	0.858	0.821	0.880
Denmark	0.792	0.860	0.769	0.737	0.851	0.839	0.843
Estonia	0.779	0.885	0.899	0.782	0.846	0.867	0.868
Finland	0.801	0.916	0.851	0.792	0.852	0.836	0.851
France	0.820	0.917	0.889	0.818	0.862	0.805	0.859
Germany	0.834	0.854	0.843	0.755	0.841	0.845	0.802
Greece	0.850	0.933	0.930	0.771	0.831	0.819	0.851
Hungary	0.823	0.929	0.912	0.778	0.872	0.844	0.878
Iceland	0.786	0.919	0.867	0.809	0.832	0.889	0.843
Ireland	0.788	0.887	0.851	0.737	0.820	0.845	0.849
Israel	0.872	0.938	0.938	0.849	0.885	0.876	0.904
Italy	0.812	0.914	0.886	0.753	0.827	0.833	0.814
Japan	0.779	0.840	0.785	0.856	0.875	0.887	0.888
Korea	0.777	0.906	0.927	0.824	0.854	0.853	0.883
Latvia	0.807	0.902	0.887	0.776	0.821	0.845	0.795
Luxembourg	0.815	0.922	0.909	0.800	0.857	0.851	0.883
Mexico	0.889	0.916	0.901	0.827	0.880	0.876	0.840
Netherlands	0.736	0.849	0.827	0.749	0.822	0.827	0.839
New Zealand	0.806	0.920	0.873	0.789	0.839	0.861	0.842
Poland	0.812	0.890	0.903	0.744	0.866	0.849	0.837
Portugal	0.850	0.943	0.911	0.806	0.866	0.859	0.859
Slovak Republic	0.840	0.923	0.903	0.801	0.867	0.861	0.843
Slovenia	0.808	0.896	0.907	0.772	0.868	0.837	0.843
Sweden	0.805	0.928	0.878	0.811	0.876	0.909	0.902
Switzerland	0.799	0.903	0.879	0.755	0.846	0.817	0.859
United Kingdom <sup>1</sup>	0.787	0.901	0.839	0.762	0.840	0.853	0.846

1. The ICT Questionnaire was only administered to a subset of students (United Kingdom excluding Scotland).

Table 16.65 Scale reliabilities for ICT Familiarity Questionnaire in partner countries and economies

	ENTUSE	HOMESCH	USESCH	INTICT	COMPCT	AUTICT	SOIAICT
B-S-J-G (China)	0.890	0.918	0.868	0.791	0.804	0.887	0.840
Brazil	0.903	0.944	0.928	0.867	0.853	0.881	0.852
Bulgaria	0.874	0.946	0.932	0.852	0.871	0.877	0.870
Colombia	0.894	0.917	0.905	0.857	0.850	0.858	0.844
Costa Rica	0.872	0.911	0.878	0.799	0.844	0.852	0.867
Croatia	0.840	0.915	0.909	0.809	0.880	0.853	0.903
Dominican Republic	0.920	0.933	0.918	0.864	0.854	0.895	0.885
Hong Kong (China)	0.842	0.931	0.930	0.800	0.843	0.913	0.895
Lithuania	0.834	0.930	0.935	0.764	0.843	0.852	0.858
Macao (China)	0.817	0.888	0.866	0.756	0.773	0.842	0.823
Peru	0.892	0.883	0.847	0.790	0.815	0.857	0.769
Russia	0.852	0.926	0.946	0.807	0.857	0.858	0.852
Singapore	0.777	0.914	0.885	0.777	0.808	0.870	0.839
Chinese Taipei	0.822	0.909	0.855	0.778	0.842	0.890	0.860
Thailand	0.888	0.929	0.924	0.848	0.850	0.869	0.821
Uruguay	0.846	0.921	0.916	0.817	0.873	0.863	0.874

## Availability and usage of ICT

Three questions in the ICT Familiarity Questionnaire asked about how often digital devices are used outside of school for leisure activities (IC008), outside of school for school work (IC010), as well as for activities in school (IC011). The answering scale for all three questions ranged from "never or hardly ever", "once or twice a month", "once or twice a week", "almost every day" to "every day". The respective indices ENTUSE (leisure activities), HOMESCH (for school work outside of

school) and USESCH (use of ICT at school) are scaled using the IRT scaling model described above. Tables 16.66, 16.67 and 16.68 shows the item wording, international item parameters and item fit for each of the three scales, respectively.

**Table 16.66 Item parameters for ICT use outside of school for leisure (ENTUSE)**

Item	How often do you use digital devices for the following activities outside of school?	Parameter estimates					
		beta	d_1	d_2	d_3	d_4	alpha
IC008Q01TA	Playing one-player games.	0.36391	-0.42410	0.57430	0.02973	-0.17992	0.62185
IC008Q02TA	Playing collaborative online games.	0.34197	-0.71734	0.48043	0.12492	0.11199	0.67610
IC008Q03TA	Using email.	0.19326	0.31538	0.26763	-0.30965	-0.27337	0.73903
IC008Q04TA	<Chatting online> (e.g. <MSN®>).	-0.22935	-0.85711	0.28840	0.30105	0.26766	0.62893
IC008Q05TA	Participating in social networks (e.g. <Facebook>, <MySpace>).	-0.41520	-0.53824	0.24012	0.18246	0.11566	0.82910
IC008Q07NA	Playing online games via social networks (e.g. <Farmville®>, <The Sims Social>).	0.50370	-0.84069	0.39629	0.13189	0.31251	0.68935
IC008Q08TA	Browsing the Internet for fun (such as watching videos, e.g. <YouTube™>).	-0.39931	0.08828	0.29156	-0.07780	-0.30204	1.44481
IC008Q09TA	Reading news on the Internet (e.g. current affairs).	-0.05522	0.09735	0.31773	-0.10172	-0.31336	1.00796
IC008Q10TA	Obtaining practical information from the Internet (e.g. locations, dates of events).	-0.02996	0.29259	0.30988	-0.17476	-0.42771	1.28358
IC008Q11TA	Downloading music, films, games or software from the internet.	-0.11231	0.33351	0.17029	-0.14709	-0.35672	1.58840
IC008Q12TA	Uploading your own created contents for sharing (e.g. music, poetry, videos, computer programs).	0.36991	-0.22438	0.22686	0.01040	-0.01287	0.92774
IC008Q13NA	Downloading new apps on a mobile device.	0.03020	0.55471	0.03540	-0.24813	-0.34198	1.56315

**Table 16.67 Item parameters for ICT use outside of school for schoolwork (HOMESCH)**

Item	How often do you use digital devices for the following activities outside of school?	Parameter estimates					
		beta	d_1	d_2	d_3	d_4	alpha
IC010Q01TA	Browsing the Internet for schoolwork (e.g. for preparing an essay or presentation).	-0.41339	1.13119	0.39950	-0.56070	-0.96999	0.79565
IC010Q02NA	Browsing the Internet to follow up lessons, e.g. for finding explanations.	-0.20642	0.78816	0.39942	-0.38277	-0.80481	0.98209
IC010Q03TA	Using email for communication with other students about schoolwork.	0.05830	0.30292	0.40577	-0.18827	-0.52043	0.94595
IC010Q04TA	Using email for communication with teachers and submission of homework or other schoolwork.	0.13185	0.58899	0.18496	-0.22650	-0.54746	1.25479
IC010Q05NA	Using social networks for communication with other students about schoolwork (e.g. <Facebook>, <MySpace>).	-0.53830	0.01470	0.63540	-0.21312	-0.43699	0.47914
IC010Q06NA	Using social networks for Communication with teachers (e.g. <Facebook>, <MySpace>).	0.17351	-0.33093	0.41087	0.01487	-0.09481	0.79062
IC010Q07TA	Downloading, uploading or browsing material from my school's website (e.g. timetable or course materials).	0.00293	0.42131	0.27136	-0.19846	-0.49422	1.06545
IC010Q08TA	Checking the school's website for announcements, e.g. absence of teachers.	0.04214	0.12897	0.34903	-0.09830	-0.37969	0.77715
IC010Q09NA	Doing homework on a computer.	-0.21896	0.70257	0.29219	-0.31999	-0.67477	0.95482
IC010Q10NA	Doing homework on a mobile device.	0.06459	0.27387	0.36208	-0.14757	-0.48838	1.02083
IC010Q11NA	Downloading learning apps on a mobile device.	0.14689	0.37227	0.20218	-0.13282	-0.44164	1.44971
IC010Q12NA	Downloading science learning apps on a mobile device.	0.21977	0.21062	0.25286	-0.07046	-0.39301	1.48379

**Table 16.68 Item parameters for Use of ICT at school in general (USESCH)**

Item	How often do you use digital devices for the following activities at school?	Parameter estimates					
		beta	d_1	d_2	d_3	d_4	alpha
IC011Q01TA	<Chatting online> at school.	-0.08101	-1.92165	1.06281	0.71262	0.14622	0.32115
IC011Q02TA	Using email at school.	0.02675	0.22049	0.42081	-0.25424	-0.38706	0.82288
IC011Q03TA	Browsing the Internet for schoolwork.	-0.40192	0.70655	0.41315	-0.39752	-0.72218	0.94650
IC011Q04TA	Downloading, uploading or browsing material from the school's website (e.g. <intranet>).	-0.05588	0.35786	0.32971	-0.18652	-0.50105	1.35374
IC011Q05TA	Posting my work on the school's website.	0.16357	0.18035	0.37225	-0.17175	-0.38085	1.36812
IC011Q06TA	Playing simulations at school.	0.23974	0.05051	0.38127	-0.08727	-0.34451	1.03355
IC011Q07TA	Practicing and drilling, such as for foreign language learning or mathematics.	0.01084	0.33078	0.42926	-0.23567	-0.52437	0.85901
IC011Q08TA	Doing homework on a school computer.	0.01316	0.39722	0.31216	-0.22927	-0.48011	1.16646
IC011Q09TA	Using school computers for group work and communication with other students.	-0.03205	0.56723	0.20722	-0.29528	-0.47917	1.12858

### Interest in ICT and perceived competence

PISA 2015 included four newly developed questions in the ICT Familiarity Questionnaire addressing students' ICT interest (IC013, INTICT), their perceived competence in ICT usage (IC014, COMPICT), their perceived autonomy related to ICT usage (IC015, AUTICT) and the degree to which ICT is a part of their daily social life (IC016, SOIAICT). All



questions used a four-point Likert answering scale ranging from "strongly disagree" to "strongly agree". Tables 16.69, 16.70, 16.71 and 16.72 shows the item wording, international item parameters and item fit for each of the four scales, respectively.

**Table 16.69 Item parameters for Students' ICT Interest (INTICT)**

Item	Thinking about your experience with digital media and digital devices: to what extent do you disagree or agree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
IC013Q01NA	I forget about time when I'm using digital devices.	0.43669	1.37049	0.36889	-1.73938	0.51239
IC013Q04NA	The Internet is a great resource for obtaining information I am interested in (e.g. news, sports, dictionary).	-0.24377	0.56601	0.70236	-1.26837	1.03630
IC013Q05NA	It is very useful to have social networks on the Internet.	-0.08135	0.81946	0.45866	-1.27812	1.30416
IC013Q11NA	I am really excited discovering new digital devices or applications.	0.22493	1.10714	0.21950	-1.32664	1.04545
IC013Q12NA	I really feel bad if no internet connection is possible.	0.38223	1.35827	-0.01344	-1.34483	0.56837
IC013Q13NA	I like using digital devices.	-0.20702	0.74116	0.52997	-1.27113	1.53333

**Table 16.70 Item parameters for Students' Perceived ICT Competence (COMPICT)**

Item	Thinking about your experience with digital media and digital devices: to what extent do you disagree or agree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
IC014Q03NA	I feel comfortable using digital devices that I am less familiar with.	0.21920	1.97262	0.55698	-2.52959	0.51025
IC014Q04NA	If my friends and relatives want to buy new digital devices or applications, I can give them advice.	0.11190	1.57835	0.40265	-1.98100	1.01112
IC014Q06NA	I feel comfortable using my digital devices at home.	-0.69950	1.01256	1.14309	-2.15565	0.67422
IC014Q08NA	When I come across problems with digital devices, I think I can solve them.	0.03847	1.55917	0.33360	-1.89277	1.38527
IC014Q09NA	If my friends and relatives have a problem with digital devices, I can help them.	0.13623	1.48395	0.33846	-1.82241	1.41915

**Table 16.71 Item parameters for Students' Perceived Autonomy related to ICT Use (AUTICT)**

Item	Thinking about your experience with digital media and digital devices: to what extent do you disagree or agree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
IC015Q02NA	If I need new software, I install it by myself.	0.31464	1.18861	0.26837	-1.45697	1.00771
IC015Q03NA	I read information about digital devices to be independent.	0.45793	1.60058	0.33888	-1.93946	0.72918
IC015Q05NA	I use digital devices as I want to use them.	-0.33563	1.26079	0.69274	-1.95353	0.92111
IC015Q07NA	If I have a problem with digital devices I start to solve it on my own.	-0.02182	1.33828	0.36100	-1.69929	1.26416
IC015Q09NA	If I need a new application, I choose it by myself.	-0.29154	1.17731	0.57501	-1.75232	1.07784

**Table 16.72 Item parameters for Students' ICT as a topic in Social Interaction (SOIAICT)**

Item	Thinking about your experience with digital media and digital devices: to what extent do you disagree or agree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
IC016Q01NA	To learn something new about digital devices, I like to talk about them with my friends.	-0.15657	1.50358	0.49371	-1.99730	0.97548
IC016Q02NA	I like to exchange solutions to problems with digital devices with others on the internet.	0.16938	1.58669	0.22627	-1.81296	1.05025
IC016Q04NA	I like to meet friends and play computer and video games with them.	0.05221	1.18775	0.45862	-1.64637	0.52396
IC016Q05NA	I like to share information about digital devices with my friends.	-0.01459	1.41803	0.28206	-1.70009	1.38479
IC016Q07NA	I learn a lot about digital media by discussing with my friends and relatives.	-0.03033	1.45972	0.37279	-1.83251	1.06553

## PARENT QUESTIONNAIRE

The Parent Questionnaire is an international option that countries can choose to implement. It addresses the parents of students participating in the PISA assessment. In PISA 2015, the Parent Questionnaire provided eight derived variables. All of them were scaled using the IRT scaling model described above. Four of these scales were mapped to the respective scales used in PISA 2006 so that trend comparison is possible. All derived variables from the Parent Questionnaire were scaled using IRT modelling.

An overview of all derived variables is shown in Table 16.73, and each will be described in the following sections.

**Table 16.73 Derived variables in the optional PISA 2015 Parent Questionnaire**

DV Name	Description	Question no.	Trend to PISA 2006	IRT scaling
PRESUPP	Child's past science activities	PA002		YES
CURSUPP	Parental current support for learning at home	PA003		YES
EMOSUPP	Parental emotional support	PA004		YES
PASCHPOL	School policies for parental involvement	PA007		YES
PQSCHOOL	Parents perceived school quality	PA007	YES	YES
PQGENSCI	Parents' view on science	PA033	YES	YES
PQENPERC	Parents concerns regarding environmental topics	PA035	YES	YES
PQENVOPT	Parents' view on future environmental topics	PA036	YES	YES

### Derived variables based on IRT Scaling

The PISA 2015 Parent Questionnaire provided data for eight scaled indices which will be presented along with the item content and parameters in the following sections. Tables 16.74 and 16.75 contain the scale reliabilities (Cronbach's Alpha coefficients) for all participating OECD and partner countries and economies, respectively.

**Table 16.74 Scale reliabilities for the Parent Questionnaire indices in OECD countries**

	PRESUPP	CURSUPP	EMOSUPP	PQSCHOOL	PASCHPOL	PQGENSCI	PQENPERC	PQENVOPT
Belgium <sup>1</sup>	0.729	0.742	0.848	0.836	0.807	0.862	0.804	0.810
Chile	0.803	0.800	0.850	0.887	0.847	0.878	0.841	0.874
France	0.731	0.752	0.817	0.849	0.802	0.846	0.802	0.821
Germany	0.742	0.749	0.777	0.823	0.819	0.838	0.805	0.749
Ireland	0.806	0.744	0.917	0.898	0.851	0.874	0.875	0.856
Italy	0.776	0.723	0.809	0.845	0.818	0.851	0.799	0.864
Korea	0.845	0.834	0.848	0.868	0.838	0.847	0.887	0.913
Luxembourg	0.759	0.768	0.818	0.845	0.830	0.856	0.863	0.867
Mexico	0.801	0.800	0.872	0.884	0.840	0.861	0.846	0.923
Portugal	0.775	0.770	0.779	0.859	0.844	0.818	0.826	0.907
Spain	0.781	0.731	0.853	0.897	0.866	0.863	0.870	0.889
United Kingdom <sup>2</sup>	0.808	0.744	0.932	0.912	0.857	0.884	0.863	0.830

1 For PRESUPP, items PA002Q07TA and PA002Q08TA were deleted by the country.

2 The Parent Questionnaire was only administered to a subset of students (Scotland).

**Table 16.75 Scale reliabilities for the Parent Questionnaire in partner countries and economies**

	PRESUPP	CURSUPP	EMOSUPP	PQSCHOOL	PASCHPOL	PQGENSCI	PQENPERC	PQENVOPT
Croatia	0.782	0.771	0.819	0.819	0.853	0.876	0.842	0.908
Dominican Republic	0.808	0.812	0.854	0.917	0.852	0.928	0.836	0.936
Georgia	0.720	0.754	0.779	0.881	0.835	0.790	0.831	0.906
Hong Kong (China)	0.829	0.831	0.781	0.826	0.820	0.898	0.876	0.885
Macao (China)	0.815	0.843	0.795	0.850	0.833	0.889	0.874	0.918
Malta	0.803	0.777	0.769	0.893	0.871	0.849	0.827	0.870

### Parental support

PISA 2015 measured parental support with three questions. PA002 retrospectively asked how frequently their child engaged in science-related learning activities at home when he or she was 10 years old and thus inquired about parents' support for science learning in the middle childhood years; examples are reading books about scientific topics or construction play. The answering categories were "very often", "regularly", "sometimes", "never" and had to be reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of parental support. The corresponding scale PRESUPP consists of all ten items of this question, some of which had been used in previous PISA cycles. Table 16.76 shows the item wording, international item parameters and item fit for PRESUPP.



Table 16.76 Item parameters for Child's past science activities (PRESUPP)

Item	Thinking back to when your child was about 10 years old, how often would your child have done these things?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
PA002Q01TA	Watched TV programmes about science	-0.33940	1.07762	-0.53619	-0.54143	1.30842
PA002Q02TA	Read books on scientific discoveries	-0.02541	0.84575	-0.44699	-0.39876	1.23082
PA002Q03TA	Watched, read or listened to science fiction	-0.29498	1.04045	-0.51632	-0.52412	0.83692
PA002Q04TA	Visited web sites about science topics	0.06720	0.66895	-0.32492	-0.34403	1.49265
PA002Q05TA	Attended a science club	0.80339	0.05692	-0.14065	0.08372	0.87210
PA002Q06NA	Construction play, e.g.-lego bricks>	-0.76810	0.98068	-0.79201	-0.18867	0.43817
PA002Q07NA	Took apart technical devices	0.05260	0.29790	-0.44784	0.14993	0.86777
PA002Q08NA	Fixed broken objects or items, e.g. broken electronic toys	0.11708	0.37036	-0.53145	0.16109	0.82646
PA002Q09NA	Experimented with a science kit, electronics kit, or chemistry set, used a microscope or telescope	0.18758	0.64835	-0.41434	-0.23401	0.99861
PA002Q10NA	Played computer games with a science content	-0.06372	0.66390	-0.41505	-0.24885	1.12808

PA003 asked about current parental support for learning at home, including both science-specific and general aspects of parental support. The corresponding scale (CURSUPP) consists of all items in that question, some of which had been used in previous PISA cycles. Answering categories ranged from "never or hardly ever", "once or twice a year", "once or twice a month", "once or twice a week", to "every day or almost every day". Table 16.77 shows the item wording, international item parameters and item fit for CURSUPP.

Table 16.77 Item parameters for Parental current support for learning at home (CURSUPP)

Item	How often do you or someone else in your home do the following things with your child?	Parameter estimates					
		beta	d_1	d_2	d_3	d_4	alpha
PA003Q01TA	Discuss how well my child is doing at school.	-0.99995	0.21467	0.56229	-0.14574	-0.63123	0.65088
PA003Q02TA	Eat <the main meal> with my child around a table.	-1.98683	-2.31373	-0.00007	1.31367	1.00013	0.34292
PA003Q03TA	Spend time just talking to my child.	-1.24493	-0.67030	0.67091	0.34271	-0.34332	0.63097
PA003Q04NA	Help my child with his/her science homework.	0.41497	-0.13300	0.60189	0.11597	-0.58486	0.90084
PA003Q05NA	Ask how my child is performing in science class.	0.02912	0.18737	0.48529	-0.06953	-0.60313	1.35385
PA003Q06NA	Obtain science-related materials (e.g., applications, software, study guides etc.) for my child.	0.54701	0.33774	0.20059	-0.14773	-0.39059	1.23451
PA003Q07NA	Discuss with my child how science is used in everyday life.	0.31372	0.39363	0.34023	-0.18121	-0.55264	1.69328
PA003Q08NA	Discuss <science related career> options with my child.	0.41746	0.38888	0.29604	-0.25126	-0.43366	1.19274

A new focus in PISA 2015 addressed the emotional support given by parents. Question PA004 included four items asking parents about their interest and support for students' school-related difficulties and achievements. Answering categories on a four-point Likert scale ranged from "strongly agree" to "strongly disagree". Table 16.78 shows the item wording, international item parameters and item fit for EMOSUPP.

Table 16.78 Item parameters for Parental emotional support (EMOSUPP)

Item	Thinking about <the last academic year>, to what extent do you agree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
PA004Q01NA	I am interested in my child's school activities.	0,12430	0,84668	1,27001	-2,11669	0,75303
PA004Q02NA	I am supportive of my child's efforts at school and his/her achievements.	0,01177	0,93360	0,89965	-1,83325	1,14243
PA004Q03NA	I support my child when he/she is facing difficulties at school.	0,07517	0,79412	0,93781	-1,73192	1,11850
PA004Q04NA	I encourage my child to be confident.	-0,19384	0,43333	1,16728	-1,60061	0,98605

### Parental involvement in school

The question addressing both parents' view on school quality and school policies for parental involvement (PA007) has been modified for each PISA cycle so far. Parents were asked how much they agreed with the statements about school policies. The response categories included "strongly agree", "agree", "disagree" and "strongly disagree". The responses had to be reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of parental involvement in school.

In PISA 2015, two derived variables were built. The scale addressing parental involvement (PASCHPOL) uses six newly developed items to measure different aspects of parental participation (PA007Q09NA, PA007Q11NA, PA007Q12NA, PA007Q13NA, PA007Q14NA, and PA007Q15NA). Table 16.79 shows the item wording, international item parameters and item fit for PASCHPOL.

**Table 16.79 Item parameters for School policies for parental involvement (PASCHPOL)**

Item	How much do you agree or disagree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
PA007Q09NA	My child's school provides an inviting atmosphere for parents to get involved.	-0.29549	1.45689	0.36785	-1.82473	1.23011
PA007Q11NA	My child's school provides effective communication between the school and families.	-0.35675	1.39322	0.40011	-1.79334	1.17906
PA007Q12NA	My child's school involves parents in the school's decision-making process.	0.00006	1.59285	0.25193	-1.84478	1.03779
PA007Q13NA	My child's school offers parent education (e.g. <courses on family literacy>) or family support programmes (e.g. <to assist with health, nutrition>).	0.75008	1.66258	0.28344	-1.94603	0.57040
PA007Q14NA	My child's school informs families about how to help students with homework and other school-related activities.	0.23771	1.38616	0.28893	-1.67509	1.09151
PA007Q15NA	My child's school cooperates with <community services> to strengthen school programmes and student development.	0.10857	1.51714	0.41933	-1.93647	0.89113

The trend indicator PQSCHOOL uses seven trend items to summarize parents' perceptions of the quality of school learning (PA007Q01TA, PATA007Q02TA, PA007Q03TA, PA007Q04TA, PA007Q05TA, PA007Q06TA, and PA007Q07TA). The same scale was used in PISA 2006, 2009, and 2012. It was scaled in such a way that a trend comparison is possible between PISA 2006 and 2015. Table 16.80 shows the item wording, international item parameters and item fit for PQSCHOOL.

**Table 16.80 Item parameters for Parents perceived school quality (PQSCHOOL)**

Item	How much do you agree or disagree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
PA007Q01TA	Most of my child's school teachers seem competent and dedicated.	-0.24354	1.59279	0.78792	-2.38072	0.89632
PA007Q02TA	Standards of achievement are high in my child's school.	-0.08680	2.34506	0.37890	-2.72396	0.61911
PA007Q03TA	I am happy with the content taught and the instructional methods used in my child's school.	0.07616	1.73442	0.50465	-2.23907	1.18881
PA007Q04TA	I am satisfied with the disciplinary atmosphere in my child's school.	-0.03327	1.50732	0.68950	-2.19682	0.76172
PA007Q05TA	My child's progress is carefully monitored by the school.	0.14735	1.68369	0.37150	-2.05519	1.16190
PA007Q06TA	My child's school provides regular and useful information on my child's progress.	0.17990	1.60999	0.37147	-1.98146	0.85679
PA007Q07TA	My child's school does a good job in educating students.	-0.07821	1.53553	0.46970	-2.00523	1.51535

### Parents' views on science and environmental topics

As in PISA 2006, the 2015 Parent Questionnaire took up the topic of parents' views on science and aspects of the environment.

Question PA033 included only trend items from 2006 and focused on parents' opinions on the importance of scientific approaches for their daily lives and society. The response categories included "strongly agree", "agree", "disagree" and "strongly disagree". The responses had to be reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of parents' view on science. The respective scale (PQGENSCI) was scaled in such a way that a trend comparison is possible between PISA 2006 and 2015. Table 16.81 shows the item wording, international item parameters and item fit for PQGENSCI.

**Table 16.81 Item parameters for Parents' view on science (PQGENSCI)**

Item	How much do you agree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
PA033Q02TA	<Broad science> is important to help us to understand the natural world.	-0.37395	1.27597	0.98893	-2.26490	1.01006
PA033Q06TA	<Broad science> is valuable to society.	-0.21465	1.45867	0.78733	-2.24599	1.12141
PA033Q07TA	<Broad science> is very relevant to me.	0.64417	2.04909	0.30847	-2.35756	1.00697
PA033Q08TA	I find that <broad science> helps me to understand the things around me.	0.37096	1.82657	0.50188	-2.32844	1.34284
PA033Q09TA	Advances in <broad science> usually bring social benefits.	-0.05992	1.79102	0.67006	-2.46108	0.71425

Question PA035 asked parents about their concerns related to current environmental topics (PQENPERC), while question PA036 asked about their optimism regarding the future trend of environmental topics (PQENVOPT). Both questions, PA035 and PA036, included trend items and some newly developed aspects regarding current environmental topics. Still, the scales were analysed to enable a trend comparison to PISA 2006.

For PA035, parents were asked to answer on a four-point Likert scale with the response options "this is a serious concern for me personally as well as others", "this is a serious concern for other people in my country but not for me personally", "this is a serious concern only for people in other countries", and "this is not a serious concern for anyone". The responses had to be reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of parents' concerns regarding environmental topics. Table 16.82 shows the item wording, international item parameters and item fit for PQENPERC.



Table 16.82 Item parameters for Parents' concerns regarding environmental topics (PQENPERC)

Item	Do you see the environmental issues below as a serious concern for yourself and/or others?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
PA035Q01TA	Air pollution	-0.39748	0.31475	-0.03886	-0.27589	1.00967
PA035Q03TA	Extinction of plants and animals	0.12756	0.59566	-0.05785	-0.53781	1.09251
PA035Q04TA	Clearing of forests for other land use	0.13784	0.79664	-0.23400	-0.56264	1.12582
PA035Q05TA	Water shortages	-0.03219	0.93264	-0.67558	-0.25706	1.07186
PA035Q06TA	Nuclear waste	0.10260	0.92366	-0.51496	-0.40870	0.97315
PA035Q07NA	Extreme weather conditions	0.12049	0.87692	-0.48587	-0.39105	1.01581
PA035Q08NA	Human contact with animal diseases	0.00784	0.68053	-0.40535	-0.27518	0.75381

For PA036, parents were asked to answer on a three-point Likert scale with the response options "improve", "stay about the same", and "get worse". The responses had to be reverse-coded so that higher WLEs and higher difficulty correspond to higher levels of parents' environmental optimism. Table 16.83 shows the item wording, international item parameters and item fit for PQENVOPT.

Table 16.83 Item parameters for Parents' view on future environmental topics (PQENVOPT)

Item	Do you think problems associated with the environmental issues below will improve or get worse over the next 20 years?	Parameter estimates			
		beta	d_1	d_2	alpha
PA036Q01TA	Air pollution	-0.01047	0.07795	-0.07795	0.85812
PA036Q03TA	Extinction of plants and animals	0.00748	0.43749	-0.43749	1.20731
PA036Q04TA	Clearing of forests for other land use	0.04940	0.31905	-0.31905	1.04442
PA036Q05TA	Water shortages	-0.00735	0.42547	-0.42547	1.20626
PA036Q06TA	Nuclear waste	-0.02245	0.41584	-0.41584	0.89242
PA036Q07NA	Extreme weather conditions	0.35808	0.51074	-0.51074	1.18526
PA036Q08NA	Human contact with animal diseases	-0.36865	0.59186	-0.59186	0.81450

## TEACHER QUESTIONNAIRES

The Teacher Questionnaire was implemented in PISA 2015 for the first time as an international option and all content was newly developed. Some questions were taken from the Teaching and Learning International Survey (TALIS) to enable comparisons and possible linkages. From the data, 20 derived variables can be analysed, nine of which were scaled using the IRT model described above. Due to the partial overlap in the two teacher questionnaires, some indices can be reported for all teachers (all indicators that are based on questions TC004 to TC026, e.g. teacher satisfaction), others only for science teachers or only for non-science teachers.

An overview of all derived variables is shown in Table 16.84, and each will be described in the following sections. Simple questionnaire indices are followed by those that are based on IRT scaling.

Table 16.84 Derived variables in the optional PISA 2015 Teacher Questionnaire

DV Name	Description	Question no.	Trend to PISA 2006	IRT scaling
EMPLSTAT	Employment Status Contract	TC004		
EMPLTIM1	Teacher Employment Time - 4 steps	TC005		
EMPLSTATd	Employment Status Contract - dichotomous	TC004		
EMPLTIM2	Teacher Employment Time - dichotomous.	TC005		
NSCHEMPL	Number of schools employed by – dichotomous	TC005		
OTT1	Originally trained teachers (wide definition).	TC013, TC014, TC015		
OTT2	Originally trained teachers (strict definition).	TC013, TC014, TC015		
NTEACH1-NTEACH11	Number of teacher educated for a specific subject (Subject was part of the Teacher education or training programme)	TC018		
STTMG1-STTMG11	Subject specific overlap between initial education and teaching the modal grade	TC018		
PROPDAT20	Proportion of professional development (Teacher reported).	TC020		
SATJOB	Satisfaction with the current job environment	TC026		YES
SATTEACH	Satisfaction with teaching profession	TC026		YES
TCEDUSHORT	Educational material shortage teachers view	TC028		YES
TCSTAFFSHORT	Staff shortage teachers view	TC028		YES
COLSCIT	Science teacher collaboration	TC031		YES
SETEACH	Self-efficacy related to teaching science content	TC033		YES
SECONT	Self-efficacy related to science content	TC034		YES
TC045Q01-TC045Q15	Content overlap between initial education and professional development	TC045		
EXCHT	Exchange and co-ordination for teaching	TC046		YES
TICLEAD	Transformational leadership teachers view	TC060		YES



## Simple questionnaire indices

### **Employment status**

Two questions in the Teacher Questionnaire were used to build five derived variables indicating various aspects of teachers' employment.

TC004 asked about employment status in terms of the contract duration (permanent/fixed-term contract for a year or less/fixed-term contract for more than 1 year), while TC005 addressed whether the teacher was in full-time or part-time employment (full-time/part-time more than 70%/part-time more than 50%/part-time 50% or less) at one or more schools.

The corresponding derived variables reflected the duration of employment, measured via TC004, a) on the original three-point scale (EMPLSTAT) and b) dichotomous, distinguishing a permanent position from fixed-term contracts (EMPLSTATd).

The data from TC005 was recoded to provide three indicators. EMPLTIM1 reflects the original four-point scale, EMPLTIM2 was recoded to reflect a dichotomous variable (full-time versus part-time), and NSCHEMPL indicates whether the teacher is employed by one or by more than one school simultaneously.

### **Teacher education**

The Teacher Questionnaire addressed a range of questions about teachers' initial education and professional development. This included a question on whether a career in the teaching profession was intended after completing ISCED 3 education (TCQ013, yes/no) and if a teacher education or training programme was completed (TC014, yes/no). TC015 asked about how the teacher qualification was received. Answering options included "standard teacher education or training programme", "in-service teacher education or training programme", "work-based teacher education or training programme", "training in another pedagogical profession" or "other".

These three questions were used to build the derived variables OTT1 (Originally trained teachers, broad definition) and OTT2 (Originally trained teachers, strict definition). The strict definition implies that a teacher had intended to be trained as a teacher from the very beginning of his or her career and has finished a "standard teacher education or training programme at a <educational institute which is eligible to educate or train teachers>". In the less strict definition, the teacher also had intended to be trained as a teacher all along and has finished any of the following three programs: either a "standard teacher education or training programme at a <educational institute which is eligible to educate or train teachers>" (option 1 in TC015), an "in-service teacher education or training programme" (option 2) or a "work-based teacher education or training programme" (option 3 in TC015).

TC018 enquired about the specific subjects that were included in the teacher's education or training programme or other professional qualification and asked if the respondents taught these subjects to the national modal grade for 15-year olds in the current school year. The derived variables NTEACH1 to NTEACH 11 reflect whether the teacher was trained to teach a certain subject. The same question is used to build the derived variables STTMG1 to STTMG11, indicating the subject-specific overlap between initial education and teaching the modal grade, i.e. whether a teacher currently teaches a certain subject combined with whether it was included in the teacher's initial training.

Participation in different professional development activities in the last 12 months was reported in TC020. This included participation in a "qualification programme", a "network of teachers focusing on professional development", "individual or collaborative research on a topic of interest", "mentoring and/or peer observation and coaching", "reading professional literature" and "engaging in informal dialogue with colleagues". The derived variable PROPD120 indicates whether a teacher took part in any of these activities in the past 12 months. It is important to note that this question is also included in TALIS 2008, but there it refers to a time frame of the past 18 months.

TC045 asked about 15 content topics that might have been included in the teachers' initial education and training and/or in professional development activities during the last 12 months. Teachers could select both if applicable. Amongst others, these included pedagogical competencies, student assessment practices and ICT skills for teaching. The derived variables TC045Q01 to TC045Q15 reflect the content overlap between initial education and professional development.



## Derived variables based on IRT Scaling

The PISA 2015 Teacher Questionnaire provided data for nine scaled indices which will be presented along with the item content and parameters in the following sections. Tables 16.85 and 16.86 contain the scale reliabilities (Cronbach's Alpha coefficients) for all participating OECD and partner countries and economies, respectively.

**Table 16.85 Scale reliabilities for Teacher Questionnaire indices in OECD countries**

	SATJOB	SATTEACH	TCEDUSHORT	TCSTAFFSHORT	COLSCIT	SETEACH	SECONT	EXCHT	TCLEAD
Australia	0.828	0.843	0.861	0.771	0.881	0.775	0.841	0.707	0.902
Chile	0.804	0.798	0.868	0.758	0.928	0.777	0.798	0.762	0.903
Czech Republic	0.836	0.792	0.814	0.681	0.868	0.681	0.704	0.724	0.889
Germany	0.781	0.813	0.857	0.673	0.863	0.690	0.776	0.701	0.856
Italy	0.775	0.797	0.877	0.684	0.870	N/A	N/A	0.738	0.877
Korea	0.831	0.773	0.883	0.745	0.885	0.861	0.769	0.804	0.921
Portugal	0.795	0.849	0.859	0.667	0.898	0.688	0.715	0.748	0.875
Spain	0.833	0.806	0.873	0.710	0.907	0.785	0.809	0.681	0.908
United States	0.841	0.852	0.835	0.802	0.894	0.733	0.834	0.743	0.919

Note: N/A indicates that the question has not been administered in the country.

**Table 16.86 Scale reliabilities for Teacher Questionnaire indices in partner countries and economies**

	SATJOB	SATTEACH	TCLEAD	TCEDUSHORT	TCSTAFFSHORT	COLSCIT	EXCHT	SETEACH	SECONT
B-S-J-G (China)	0.856	0.653	0.922	0.929	0.910	0.921	0.851	0.816	0.807
Brazil	0.804	0.780	0.910	0.898	0.817	0.907	0.806	0.782	0.759
Colombia	0.838	0.761	0.928	0.885	0.752	0.912	0.782	0.702	0.727
Dominican Republic	0.847	0.666	0.901	0.821	0.759	0.884	0.736	0.539	0.722
Hong Kong (China)	0.805	0.697	0.903	0.866	0.783	0.862	0.781	0.732	0.786
Macao (China)	0.804	0.801	0.884	0.863	0.839	0.886	0.756	0.785	0.834
Malaysia	0.812	0.764	0.926	0.908	0.827	0.889	0.819	0.747	0.823
Peru	0.808	0.733	0.905	0.875	0.783	0.897	0.776	0.759	0.799
Chinese Taipei	0.858	0.761	0.921	0.896	0.768	0.886	0.824	0.795	0.785
United Arab Emirates	0.823	0.788	0.919	0.919	0.863	0.916	0.750	0.783	0.764

## Job satisfaction and school leadership

The teacher questionnaires used one question (TC026) to ask about teachers' job satisfaction. The four-point Likert scale ranged from "strongly agree", "agree", "disagree" to "strongly disagree". The derived variable "satisfaction with the current job environment" (SATJOB) was scaled using items TC026Q05NA, TC026Q07NA, TC026Q09NA, TC026Q10NA. The derived variable "satisfaction with teaching profession" (SATTEACH) was scaled using items TC026Q01NA, TC026Q02NA, TC026Q04NA (recoded), and TC026Q06N (recoded). Tables 16.87 and 16.88 show the item wording, international item parameters and item fit for SATJOB and SATTEACH, respectively.

**Table 16.87 Item parameters for Satisfaction with the current job environment (SATJOB)**

Item	We would like to know how you generally feel about your job. How strongly do you agree or disagree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
TC026Q05NA	I enjoy working at this school.	0.08511	1.62541	0.54615	-2.17156	1.25762
TC026Q07NA	I would recommend my school as a good place to work.	0.36952	1.70996	0.43825	-2.14821	1.13821
TC026Q09NA	I am satisfied with my performance in this school.	-0.33945	1.85826	0.92920	-2.78746	0.74091
TC026Q10NA	All in all, I am satisfied with my job.	-0.31986	1.81916	0.78320	-2.60236	0.86326

**Table 16.88 Item parameters for Satisfaction with teaching profession (SATTEACH)**

Item	We would like to know how you generally feel about your job. How strongly do you agree or disagree with the following statements?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
TC026Q01NA	The advantages of being a teacher clearly outweigh the disadvantages.	0.00964	1.50124	0.58113	-2.08236	0.57877
TC026Q02NA	If I could decide again, I would still choose to work as a teacher.	0.08215	1.14089	0.25229	-1.39318	1.33432
TC026Q04NA	I regret that I decided to become a teacher.	-0.33769	0.93611	0.39061	-1.32672	1.25278
TC026Q06NA	I wonder whether it would have been better to choose another profession.	0.36908	1.63123	-0.20855	-1.42269	0.83412

TC060 asked about teachers' views on school leadership (TCLEAD). The items can be related to those used in SC009. The four-point Likert scale ranged from "strongly agree", "agree", "disagree" to "strongly disagree". Table 16.89 shows the item wording, international item parameters and item fit for TCLEAD.

**Table 16.89 Item parameters for Transformational leadership teachers view (TCLEAD)**

Item	To what extent do you disagree or agree with the following statements regarding your school?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
TC060Q02NA	The principal tries to achieve consensus with all staff when defining priorities and goals in school.	-0.10263	1.86481	0.49417	-2.35898	1.04441
TC060Q04NA	The principal is aware of my needs.	0.13401	2.15078	0.39237	-2.54316	1.05004
TC060Q06NA	The principal inspires new ideas for my professional learning.	0.25823	2.21824	0.34937	-2.56761	0.75192
TC060Q07NA	The principal treats teaching staff as professionals.	-0.62260	1.59634	0.73520	-2.33154	0.92705
TC060Q09NA	The principal ensures our involvement in decision making.	0.28493	1.95661	0.30381	-2.26042	1.22657

### Educational resources

In parallel to the questions addressing shortage of educational resources in the School Questionnaire (SC017), teachers were asked whether their school's capacity to provide instruction is hindered (TC028) due to lack of educational resources (TCEDUSHORT) or staff shortage (TCSTAFFSHORT). The four-point Likert scales ranged from "not at all", "very little", to "to some extent", and "a lot". The respective IRT scaled derived variables used items TC028Q05NA, TC028Q06NA, TC028Q07NA, TC028Q08NA (TCEDUSHORT) and TC028Q01NA, TC028Q02NA, TC028Q03NA, TC028Q04NA (TCSTAFFSHORT). Tables 16.90 and 16.91 show the item wording, international item parameters and item fit for TCEDUSHORT and TCSTAFFSHORT, respectively.

**Table 16.90 Item parameters for Educational material shortage teachers view (TCEDUSHORT)**

Item	Is your school's capacity to provide instruction hindered by any of the following issues?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
TC028Q05NA	A lack of educational material (e.g. textbooks, IT equipment, library or laboratory material).	-0.00584	1.79252	0.20125	-1.99377	0.38660
TC028Q06NA	Inadequate or poor quality educational material (e.g. textbooks, IT equipment, library or laboratory material).	0.17238	1.96513	0.13631	-2.10144	0.41894
TC028Q07NA	A lack of physical infrastructure (e.g. building, grounds, heating/cooling, lighting and acoustic systems).	-0.02702	1.46428	0.05783	-1.52211	1.60609
TC028Q08NA	Inadequate or poor quality physical infrastructure (e.g. building, grounds, heating/cooling, lighting and acoustic systems).	-0.01673	1.54394	0.03828	-1.58222	1.58837

**Table 16.91 Item parameters for Staff shortage teachers view (TCSTAFFSHORT)**

Item	Is your school's capacity to provide instruction hindered by any of the following issues?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
TC028Q01NA	A lack of teaching staff.	0.05865	0.82239	0.28125	-1.10364	0.70148
TC028Q02NA	Inadequate or poorly qualified teaching staff.	0.27045	1.09231	-0.02139	-1.07092	0.94076
TC028Q03NA	A lack of assisting staff.	-0.36633	0.77807	0.15947	-0.93754	0.97423
TC028Q04NA	Inadequate or poorly qualified assisting staff.	0.04432	0.90031	0.00200	-0.90231	1.38352

### Teaching and teacher collaboration

Science teacher collaboration (COLSCIT) was assessed asking about teachers' agreement on a four-point Likert scale ranging from "strongly disagree" to "strongly agree" regarding different aspects of cooperation (SC031). Table 16.92 shows the item wording, international item parameters and item fit for COLSCIT.

**Table 16.92 Item parameters for Science teacher collaboration (COLSCIT)**

Item	To what extent do you disagree or agree with the following statements about regular cooperation among your fellow <school science> teachers and yourself?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
TC031Q04NA	We discuss the achievement requirements for <school science> when setting tests.	-0.16011	1.28360	0.56727	-1.85087	1.05515
TC031Q07NA	It is natural for us to cooperate on what homework to give to our students.	0.40101	1.59176	0.30721	-1.89897	0.87181
TC031Q11NA	We discuss the criteria we use to grade written tests.	-0.26537	1.30104	0.47756	-1.77860	0.97670
TC031Q13NA	We exchange tasks for lessons and homework that cover a range of different levels of difficulty.	0.12939	1.54771	0.32741	-1.87513	0.96791
TC031Q14NA	I prepare a selection of teaching units with my fellow <school science> teachers.	0.22109	1.48868	0.28867	-1.77735	0.98267
TC031Q15NA	We discuss ways to teach learning strategies and techniques to our students.	-0.14758	1.39740	0.46003	-1.85743	1.22837
TC031Q18NA	My fellow <school science> teachers benefit from my specific skills and interests.	0.02407	1.63974	0.56793	-2.20767	0.89188
TC031Q20NA	We discuss ways to better identify students' individual strengths and weaknesses.	-0.10157	1.54729	0.39715	-1.94445	1.02551



TC046 addressed teaching-related co-operation using items like “teaching jointly” or “exchanging teaching materials”. Teachers were asked to rate these activities with the following answering categories “never”, “once a year or less”, “2-4 times a year”, “5-10 times a year”, “1-3 times a month”, and “once a week or more”. The derived variable indicates exchange and co-ordination for teaching (EXCHT, items TC046Q04NA, TC046Q05NA, TC046Q06NA, TC046Q07NA). Table 16.93 shows the item wording, international item parameters and item fit for EXCHT.

**Table 16.93 Item parameters for Exchange and co-ordination for teaching (EXCHT)**

Item	On average, how often do you do the following in this school?	Parameter estimates						
		beta	d_1	d_2	d_3	d_4	d_5	alpha
TC046Q04NA	Exchange teaching materials with colleagues	0.01092	0.72353	0.55623	-0.32800	-0.27921	-0.67255	0.79876
TC046Q05NA	Engage in discussions about the learning development of specific students	-0.07710	0.71516	0.42550	-0.26368	-0.24930	-0.62768	1.34674
TC046Q06NA	Work with other teachers in my school to ensure common standards in evaluations for assessing student progress	0.14917	0.69949	0.50130	-0.25550	-0.21187	-0.73343	1.27682
TC046Q07NA	Attend team conferences	-0.16507	0.65883	0.83414	-0.08830	-0.46112	-0.94355	0.57768

The Teacher Questionnaire also addressed teachers' self-efficacy related to teaching science content (SETEACH) such as using experiments in everyday teaching (TC033) and self-efficacy related to science content (SECONT) such as explaining a complex scientific concept to a fellow teacher (TC034). Teachers were asked to rate their agreement with different statements on a four-point Likert scale with the answering options “not at all”, “very little”, “to some extent”, “to a large extent”. Tables 16.94 and 16.95 show the item wording, international item parameters and item fit for SETEACH and SECONT, respectively.

**Table 16.94 Item parameters for Self-efficacy related to teaching science content (SETEACH)**

Item	To what extent can (or could) you do the following?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
TC033Q04NA	Design experiments and hands-on activities for <inquiry-based learning>	0.12659	1.83102	0.16742	-1.99844	0.78105
TC033Q05NA	Assign tailored tasks to the weakest as well as to the best students	0.22470	1.70739	0.20821	-1.91560	0.99481
TC033Q06NA	Use a variety of assessment strategies	-0.10617	1.62195	0.21158	-1.83353	1.15511
TC033Q08NA	Facilitate a discussion among students on how to interpret experimental findings	-0.18687	1.56568	0.29511	-1.86079	1.06903

**Table 16.95 Item parameters for Self-efficacy related to science content (SECONT)**

Item	To what extent can (or could) you do the following?	Parameter estimates				
		beta	d_1	d_2	d_3	alpha
TC034Q01NA	Explain a complex scientific concept to a fellow teacher	-0.07070	1.84108	0.14818	-1.98925	1.02075
TC034Q02NA	State and defend an informed position on ethical problems relating to <broad science>	0.11600	1.73719	0.15930	-1.89649	1.21705
TC034Q04NA	Read state-of-the art papers in my scientific discipline	0.08670	1.65776	0.06347	-1.72124	0.86318
TC034Q06NA	Explain the links between biology, physics and chemistry	-0.16001	1.67677	0.24832	-1.92509	0.89902

## THE PISA INDEX OF ECONOMIC, SOCIAL AND CULTURAL STATUS (ESCS)

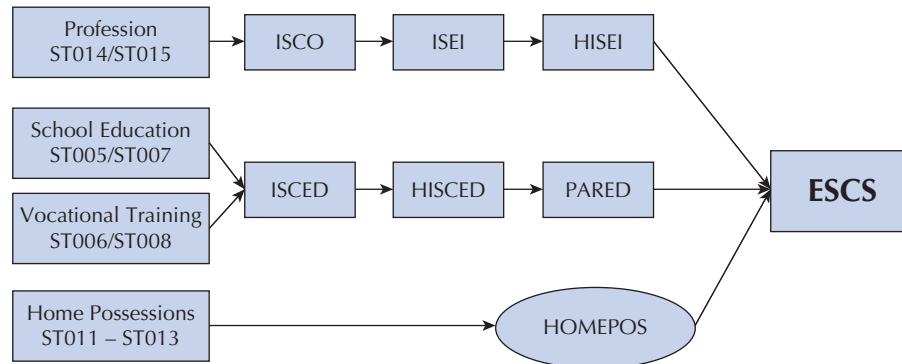
### Computation of ESCS

The ESCS is a composite score built by the indicators parental education (PARED), highest parental occupation (HISEI), and home possessions (HOMEPOS) including books in the home via principal component analysis (PCA). (See description of these three variables above). The rationale for using these three components was that socio-economic status has usually been seen as based on education, occupational status and income. As no direct income measure has been available from the PISA data, the existence of household items has been used as a proxy for family wealth.

For students with missing data on one out of the three components, the missing variable was imputed. Regression on the other two variables was used to predict the third (missing) variable, and a random component was added to the predicted value. If there were missing data on more than one component, ESCS was not computed and a missing value was assigned for ESCS. After imputation, all three components were standardised for OECD countries<sup>7</sup> and partner countries/economies with an OECD mean of zero and a standard deviation of one.

Standardised variables, including imputed values, were used in the PCA to obtain ESCS values. As in previous cycles, ESCS was defined as the component score for the first principal component. The PCA was run across equally weighted countries, including OECD as well as partner countries/economies.

■ Figure 16.5 ■

**Computation of ESCS in PISA 2015**

Note: ISCO: International Standard Classification of Occupations; ISEI: occupational status of mother and father; HISEI: highest parental occupational status; ISCED: International Standard Classification of Education; HISCED: Highest education of parents (ISCED); PARED: Index for highest parental education in years of schooling; HOMEPOS: Index of home possessions (WLE); ESCS: Index of economic, social and cultural status.

Please note that in previous cycles, the PCA was based on OECD countries only. For partner countries/economies, ESCS scores were simple indices using standardised imputed variables, fixed factor scores from PCA across OECD countries, and the eigenvalue of the first principal component (please see PISA 2012 Technical Report<sup>8</sup>). In PISA 2015, the PCA is estimated across all OECD and partner countries/economies concurrently<sup>9</sup>. Thus, all countries and economies contribute equally to the estimation of ESCS scores. However, for the purpose of reporting the ESCS scale has been transformed with zero being the score of an average OECD student and one being the standard deviation across equally weighted OECD countries<sup>10</sup>.

**Consistency across countries**

Using principal component analysis (PCA) to derive factor loadings for each participating country provided insight into the extent to which relationships of the index were similar between the three variables. Table 16.96 shows the PCA results for the OECD countries and Table 16.97 shows those for partner countries/ economies. The tables also include the scale reliabilities (Cronbach's alpha) for the z-standardised variables.

**Table 16.96 Factor loadings and reliability (Cronbach's Alpha) of ESCS 2015 in OECD countries**

	HISEI	PARED	HOMEPOS	Reliability
Australia	0.80	0.79	0.67	0.60
Austria	0.81	0.79	0.72	0.66
Belgium	0.84	0.79	0.71	0.68
Canada	0.80	0.79	0.64	0.58
Chile	0.85	0.84	0.77	0.76
Czech Republic	0.82	0.76	0.72	0.65
Denmark	0.83	0.79	0.68	0.65
Estonia	0.83	0.78	0.68	0.63
Finland	0.80	0.76	0.68	0.59
France	0.83	0.78	0.72	0.66
Germany	0.83	0.81	0.74	0.70
Greece	0.83	0.82	0.71	0.70
Hungary	0.85	0.83	0.75	0.74
Iceland	0.75	0.76	0.65	0.53
Ireland	0.81	0.80	0.70	0.65
Israel	0.80	0.79	0.68	0.60
Italy	0.83	0.79	0.72	0.68
Japan	0.74	0.76	0.68	0.54
Korea	0.78	0.79	0.73	0.62
Latvia	0.83	0.82	0.72	0.69
Luxembourg	0.86	0.79	0.75	0.72
Mexico	0.85	0.85	0.80	0.77
Netherlands	0.81	0.78	0.75	0.67
New Zealand	0.81	0.75	0.68	0.58
Norway	0.80	0.78	0.68	0.60
Poland	0.81	0.80	0.71	0.65
Portugal	0.86	0.84	0.76	0.75
Slovak Republic	0.84	0.82	0.74	0.72
Slovenia	0.84	0.82	0.69	0.68
Spain	0.85	0.83	0.74	0.73
Sweden	0.82	0.77	0.66	0.61
Switzerland	0.82	0.81	0.69	0.68
Turkey	0.82	0.79	0.77	0.68
United Kingdom	0.80	0.76	0.73	0.63
United States	0.84	0.81	0.74	0.71



Table 16.97 Factor loadings and reliability (Cronbach's Alpha) of ESCS 2015 in partner countries and economies

	HISEI	PARED	HOMEPOS	Reliability
Albania	0.82	0.82	0.74	0.69
Algeria	0.79	0.76	0.72	0.62
Argentina	0.84	0.82	0.75	0.72
B-S-J-G (China)	0.84	0.80	0.80	0.74
Brazil	0.82	0.80	0.78	0.71
Bulgaria	0.82	0.81	0.69	0.67
Colombia	0.81	0.78	0.80	0.70
Costa Rica	0.82	0.79	0.82	0.73
Croatia	0.82	0.80	0.70	0.67
Cyprus*	0.85	0.82	0.70	0.70
Dominican Republic	0.79	0.77	0.75	0.66
FYROM	0.77	0.76	0.72	0.61
Georgia	0.78	0.76	0.76	0.62
Hong Kong (China)	0.84	0.81	0.76	0.73
Indonesia	0.83	0.81	0.80	0.74
Jordan	0.81	0.83	0.73	0.67
Kazakhstan	0.72	0.77	0.69	0.44
Kosovo	0.76	0.76	0.70	0.58
Lebanon	0.60	0.79	0.77	0.54
Lithuania	0.83	0.81	0.71	0.68
Macao (China)	0.79	0.80	0.70	0.64
Malaysia	0.85	0.76	0.80	0.73
Malta	0.84	0.82	0.65	0.67
Moldova	0.80	0.76	0.76	0.65
Montenegro	0.79	0.76	0.70	0.61
Peru	0.86	0.82	0.81	0.76
Qatar	0.74	0.78	0.50	0.38
Romania	0.81	0.78	0.74	0.67
Russia	0.80	0.80	0.70	0.63
Singapore	0.83	0.82	0.77	0.73
Chinese Taipei	0.79	0.80	0.75	0.67
Thailand	0.82	0.81	0.79	0.72
Trinidad and Tobago	0.76	0.75	0.70	0.57
Tunisia	0.83	0.79	0.80	0.73
United Arab Emirates	0.74	0.79	0.48	0.36
Uruguay	0.83	0.82	0.77	0.73
Viet Nam	0.82	0.83	0.80	0.74

\* See note under Table 16.6.

## Trends in ESCS

### ESCS model

The index of ESCS was used first in the PISA 2000 analysis and at that time was derived from five indices: highest occupational status of parents (HISEI), highest educational level of parents (PARED), and three IRT scales based on student reports on home possessions: family wealth (WEALTH), cultural possessions (CULTPOSS) and home educational resources (HEDRES).

Since PISA 2003 the ESCS is derived from three indices: highest parental occupation (HISEI), highest parental education (PARED), and one IRT scale based on student reports on home possessions including books in the home (HOMEPOS). However, until PISA 2012 the PCA was based on OECD countries only. In PISA 2015, the PCA is estimated across all countries concurrently. Thus, all countries and economies contribute equally to the estimation of ESCS scores.

### ESCS components

The mapping of ISCED levels to years of schooling (PARED) was updated in 2009 and 2015 for some countries, taking into account changes in countries' educational systems.

Indicators of HOMEPOS have been dropped or added in all PISA cycles (except in PISA 2012) taking into account social, technical and economic changes in participating societies. Moreover, the method for HOMEPOS estimation has changed in PISA 2009, PISA 2012 and PISA 2015.

Since PISA 2012 parental occupation is coded into HISEI using the current international standard classification of occupations, ISCO-08. Previous cycles used ISCO-88. For the effects of ISCO-08 compared to ISCO-88 on ESCS and performance please see PISA 2012 Technical Report, pp. 372 (OECD, 2014).

In conclusion, ESCS components and the ESCS model has changed over cycles and with that, ESCS scores are not comparable across cycles directly. In order to enable a trends study, in PISA 2015 the ESCS was computed for the current cycle and also recomputed for the earlier cycles using a similar methodology.

### **ESCS trend scores**

Before trend scores could be estimated, slight adjustments of the three trend components had to be made. As in PISA 2012 the occupational coding scheme involved in the process of forming HISEI changed from ISCO-88 to ISCO-08, the occupational codes for previous cycles were mapped from the former to the current scheme (see also PISA 2012 Technical Report, Chapter 3 (OECD, 2014)).

In order to make the PARED component comparable across cycles, similar ISCED to PARED mapping schemes were employed for all the cycles. These mappings to years of education can be found in Annex E. To make the HOMEPOS component more comparable across cycles, the variable *books in the home* (ST013Q01TA) was recoded into a four-level categorical variable (fewer than or equal to 25 books, 26-100 books, 101-500 books, more than 500 books).

The HOMEPOS scale was constructed in three steps. In the first step, international item parameters for all items (except country-specific items, i.e. ST011Q17NA, ST011Q18NA and ST011Q19NA) administered in PISA 2015 were obtained from a concurrent calibration of the 2015 data. Except for the recoding of variable ST013Q01TA, this step is identical with the regular scaling of HOMEPOS in PISA 2015 (see above). In the second step, items from all previous cycles (i.e., 2000-2012) were scaled whereas parameters were fixed for all items administered in 2015 and for which no unique (i.e., country-specific) item parameters became necessary (see Table 16.9 for the respective subset of items and their parameters). Item parameters for all other items (except national items) were freely estimated but constrained to be equal across countries within cycles. Only national items (i.e., ST011Q17NA, ST011Q18NA and ST011Q19NA) received unique parameters throughout. Additional analyses on the invariance of item parameters across countries, languages and cycles were conducted and unique parameters were assigned if necessary. Once this process was finished, WLEs for all students from previous cycles (2000-2012) were estimated in the third and final step. By restricting the largest subset of items (17 out of 27) to be equal across cycles, the HOMEPOS scores can be regarded to be on a joint scale, allowing for comparisons of countries across cycles and thus allowing to be used in the calculation of trend ESCS.

The PCA for obtaining ESCS scores was then calculated as described in the section “ESCS computation” above. However, the calculation was done across all cycles using these three comparable components (HISEI, PARED, and HOMEPOS).



## Notes

1. For ease of understanding, the scaling constant, D, has been omitted from formulas 16.1 to 16.3 (refer chapter 9 for details).
2. For standardisation, data were grouped by national centre rather than by country; as a result, data for the United Kingdom (GBR) comprised two sets QUK (United Kingdom excluding Scotland) and QSC (Scotland) and data for Belgium comprised two sets, BFL (Flemish Community) and BFR (French- and German-speaking Community), thus contributing as OECD countries with double weight each.
3. Based on pseudo counts from the E-step (during the EM algorithm).
4. It should be noted that research on the validity of this procedure is still ongoing. Further empirical evidence is needed to support setting the cut-off value of .3 and its implications, meanwhile this approach can be compared with other psychometric methods to evaluate cross-cultural comparability (e.g. He and Kubacka, 2015).
5. See Annex D.
6. <https://stats.oecd.org/glossary/detail.asp?ID=1436>
7. In line with the standardisation of the IRT-based Derived Variables, the United Kingdom (GBR) and Belgium contributed two samples each, with each sample's weight equal to that of other countries.
8. <https://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm>
9. For Spain, the QES sample was included in the Principal Component Analysis whereas for ESCS standardization, ESP was used to compute the OECD transformation constants.
10. In October 2016, it turned out that the PARED variable was coded incorrectly for Spain, Lebanon and Latvia. As a consequence, the ESCS calculation was based on incorrect variables in some of the countries. To avoid changing the values of ESCS for all countries, at a time where most reports were already completed, ESCS was recalculated only for the samples with mistakes in the original PARED values, using the results from the international ESCS calculation (i.e., constants for standardizing input variables, factor loadings, eigenvalue, and constants for standardising the ESCS). As a consequence of this partial recalculation, the ESCS mean across OECD countries is no longer exactly zero and the standard deviation is no longer exactly one. Instead, the respective descriptives are -0.0259 and 1.00001.

## References

- Bempechat, J., N. V. Jimenez and B. A. Boulay (2002), "Cultural-cognitive issues in academic achievement: New directions for cross-national research", in A. C. Porter and A. Gamoran (eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement*, National Academic Press, Washington, D.C.
- Buckley, J. (2009), "Cross-national response styles in international educational assessments: Evidence from PISA 2006", [https://edsurveys.rti.org/PISA/documents/Buckley\\_PISAreponsestyle.pdf](https://edsurveys.rti.org/PISA/documents/Buckley_PISAreponsestyle.pdf) (accessed on November 2 2016).
- Ganzeboom, H.B.G. and D.J. Treiman (2003), "Three internationally standardised measures for comparative research on occupational status", in J.H.P. Hoffmeyer-Zlotnik and C. Wolf (eds.), *Advances in Cross-National Comparison, A European Working Book for Demographic and Socio-Economic Variables*, pp. 159-193, Kluwer Academic Press, New York.
- 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: Background, Technical Issues and Methods of Data Analysis*, pp. 97-115, Springer, New York.
- He, J. and K. Kubacka (2015), "Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013", *OECD Education Working Papers*, No. 124, OECD Publishing, Paris.
- International Labour Office (2012), International standard classification of occupations, ISCO-08, [www.ilo.org/public/english/bureau/stat/isco/](http://www.ilo.org/public/english/bureau/stat/isco/) (accessed on November 2016).
- Masters, G. N. and B. D. Wright (1997), "The partial credit model", in W. J. van der Linden and R. K. Hambleton (eds.), *Handbook of Modern Item Response Theory*, Springer, New York.
- Muraki, E. (1992), "A generalized partial credit model: Application of an EM algorithm", *ETS Research Report Series*, Vol. 1992/1, pp. i-30.
- OECD (2017), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*, PISA, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264281820-en>.

Penfield, R. D., N. D. Myers and E. W. Wolfe: OECD (2014), *PISA 2012 Technical Report*, PISA, OECD publishing, Paris.

OECD (1999), *Classifying Educational Programmes: Manual for ISCED-97 Implementation in OECD Countries*, OECD Publishing, Paris, <http://www.oecd.org/education/skills-beyond-school/1962350.pdf>.

Penfield, R. D., N. D. Myers and E. W. Wolfe (2008), "Methods for Assessing Item, Step, and Threshold Invariance in Polytomous Items Following the Partial Credit Model", *Educational and Psychological Measurement*, Vol. 68, pp. 717-733.

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Nielsen and Lydiche, Copenhagen, Denmark.

von Davier, M. (2008), "A general diagnostic model applied to language testing data", *British Journal of Mathematical and Statistical Psychology*, Vol. 61, pp. 287-307.

Warm, T. A. (1989), Weighted likelihood estimation of ability in item response theory, *Psychometrika*, Vol. 54, pp. 427-450.



17

# Questionnaire design and computer-based questionnaire platform

<b>Introduction .....</b>	346
<b>General questionnaire process.....</b>	346
<b>Step 1: Master questionnaires design .....</b>	347
<b>Step 2: Master questionnaires authoring.....</b>	350
<b>Step 3: Creation of national questionnaires .....</b>	362
<b>Step 4: National questionnaire adaptation and translation .....</b>	362
<b>Step 5: National questionnaires quality check.....</b>	363
<b>Step 6: Preparation of national questionnaires for delivery.....</b>	363
<b>Step 7: Data collection and quality monitoring .....</b>	365
<b>Step 8: Completion of data collection.....</b>	367
<b>Development process overview and technical infrastructure .....</b>	367
<b>Conclusion.....</b>	367

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.



## INTRODUCTION

Questionnaires have been important components of the PISA survey from its beginning. They have gained substantially in importance by delivering information about the learning contexts in countries and providing standalone reporting indicators in addition to merely explaining the “background” for reporting cognitive test results. The format and design of the questionnaires have changed across the different PISA cycles and the transition from paper-based to computer-based assessment began slowly for the questionnaires instruments since PISA 2012. While optional online administration for the School Questionnaire was already introduced in PISA 2012, PISA 2015 provided all questionnaires on computer.

As shown in Table 17.1, a number of questionnaires, both compulsory and optional, were implemented in PISA 2015.

**Table 17.1 PISA 2015 questionnaires**

Questionnaire	Mode of delivery	Compulsory
Student Questionnaire	Computer and paper	Yes
School Questionnaire	Computer and paper	Yes
Educational Career Questionnaire	Computer only	No
ICT Questionnaire	Computer only	No
Teacher Questionnaire	Computer only	No
Parent Questionnaire	Paper only	No

Computer-based delivery was the standard administration format, with the exception of the parent questionnaire option for all countries that implemented it, and a minority of countries that still used paper-based delivery for all tests and questionnaires. The student questionnaires were delivered as part of the student delivery platform and presented on the schools’ computers. The School Questionnaire and the optional Teacher Questionnaires were administered online. The electronic assessment allowed for several types of innovations but the major purpose was to increase the data quality and the response rate for this study.

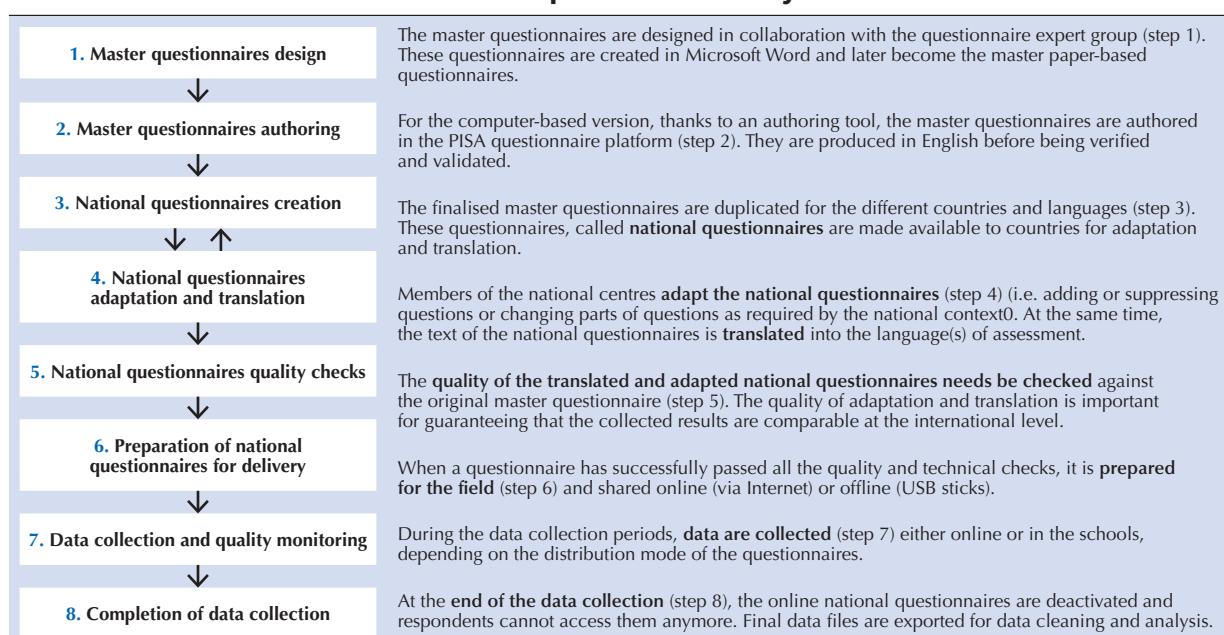
After providing a global overview of the questionnaire implementation process used for PISA 2015, this chapter explains the PISA 2015 design for both the paper-based and the computer-based questionnaires in the field trial and the main survey. The next sections describe the computer-based questionnaires, the PISA questionnaire platform and its functionalities.

## GENERAL QUESTIONNAIRE PROCESS

The questionnaire life cycle in PISA follows a process that can be split in eight major steps described in Figure 17.1.

■ Figure 17.1 ■

### PISA 2015 questionnaire life cycle





For each cycle of the PISA survey, this sequence of steps takes place twice: once for the field trial and once for the main survey. During the field trial, the whole platform (i.e. the tools, computer servers, network access, etc.) and the material (i.e. the questionnaires) are tested on a limited sample of respondents. Between the field trial and the main survey, the collected results and feedback are analysed. Then, for the main survey, the sequence is started for a second time and each step integrates all necessary adjustments in terms of process, questionnaires material, and tooling. This double-phase cycle provides better data quality.

In the following sections, each step of this process is explained in more detail.

## STEP 1: MASTER QUESTIONNAIRES DESIGN

Starting with the first cycle in 2000, PISA has emphasised the importance of collecting context information from students and schools along with the assessment of student achievement. A Student Questionnaire (StQ – approximately 30 minutes) and a School Questionnaire (ScQ – approximately 45 minutes) cover a broad range of contextual variables. The content of these questionnaires – especially the content of the StQ – has changed considerably between cycles, but the design has remained stable: every student participating in the PISA assessment completes the StQ, and every school principal (one per school) completes the ScQ. (Please also see Chapter 3 about the context questionnaire development).

PISA has also included several international options, i.e. additional instruments that countries could administer on a voluntary basis. For PISA 2015, it included a Parent Questionnaire (PAQ) as well as optional questionnaires for the students including the Educational Career Questionnaire (ECQ) and ICT Familiarity Questionnaire (ICTQ). In addition, for the first time, PISA 2015 included a Teacher Questionnaire (TCQ) as an international option into its design. Table 17.2 summarises the participation of countries/economies in the international questionnaires.

[Part 1/2]

Table 17.2 Questionnaire participation in PISA 2015 main survey

Country/economy	Mode	Student	School	Ed. Career	ICT	Student UH	Teacher	Parent
<b>OECD</b>								
Australia	CBA	Yes	Yes	Yes	Yes		Yes	
Austria	CBA	Yes	Yes		Yes	Yes		
Belgium	CBA	Yes	Yes	Yes	Yes	Yes		Yes
Canada	CBA	Yes	Yes					
Chile	CBA	Yes	Yes		Yes		Yes	Yes
Czech Republic	CBA	Yes	Yes		Yes	Yes	Yes	
Denmark	CBA	Yes	Yes		Yes	Yes		
Estonia	CBA	Yes	Yes		Yes			
Finland	CBA	Yes	Yes		Yes	Yes		
France	CBA	Yes	Yes		Yes			Yes
Germany	CBA	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Greece	CBA	Yes	Yes	Yes	Yes			
Hungary	CBA	Yes	Yes	Yes	Yes			
Iceland	CBA	Yes	Yes	Yes	Yes			
Ireland	CBA	Yes	Yes		Yes			Yes
Israel	CBA	Yes	Yes		Yes			
Italy	CBA	Yes	Yes	Yes	Yes		Yes	Yes
Japan	CBA	Yes	Yes		Yes			
Korea	CBA	Yes	Yes	Yes	Yes		Yes	Yes
Latvia	CBA	Yes	Yes	Yes	Yes			
Luxembourg	CBA	Yes	Yes		Yes			Yes
Mexico	CBA	Yes	Yes		Yes			Yes
Netherlands	CBA	Yes	Yes		Yes	Yes		
New Zealand	CBA	Yes	Yes		Yes			
Norway	CBA	Yes	Yes					
Poland	CBA	Yes	Yes	Yes	Yes			
Portugal	CBA	Yes	Yes		Yes		Yes	Yes
Slovak Republic	CBA	Yes	Yes	Yes	Yes	Yes		
Slovenia	CBA	Yes	Yes	Yes	Yes	Yes		
Spain	CBA	Yes	Yes	Yes	Yes		Yes	Yes
Sweden	CBA	Yes	Yes		Yes			
Switzerland	CBA	Yes	Yes		Yes			
Turkey	CBA	Yes	Yes					
United Kingdom (excluding Scotland)	CBA	Yes	Yes		Yes			Yes
United Kingdom (Scotland)	CBA	Yes	Yes					Yes
United States	CBA	Yes	Yes	Yes		Yes	Yes	
United States (Puerto Rico)	PBA	Yes	Yes					

Note: CBA = Computer-Based Assessment, PBA = Paper-Based Assessment. UH = “Une heure” shortened questionnaire version.

[Part 2/2]

Table 17.2 Questionnaire participation in PISA 2015 main survey)

Country/economy	Mode	Student	School	Ed. Career	ICT	Student UH	Teacher	Parent
<b>PARTNER</b>								
Albania	PBA	Yes	Yes					
Algeria	PBA	Yes	Yes					
Argentina	PBA	Yes	Yes					
Brazil	CBA	Yes	Yes		Yes		Yes	
B-S-J-G (China)**	CBA	Yes	Yes	Yes	Yes		Yes	
Bulgaria	CBA	Yes	Yes	Yes	Yes			
Colombia	CBA	Yes	Yes		Yes		Yes	
Costa Rica	CBA	Yes	Yes		Yes	Yes		
Croatia	CBA	Yes	Yes	Yes	Yes			Yes
Cyprus*	CBA	Yes	Yes					
Dominican Republic	CBA	Yes	Yes		Yes		Yes	Yes
FYROM	PBA	Yes	Yes					
Georgia	PBA	Yes	Yes					Yes
Hong Kong (China)	CBA	Yes	Yes	Yes	Yes		Yes	Yes
Indonesia	PBA	Yes	Yes					
Jordan	PBA	Yes	Yes					
Kazakhstan	PBA	Yes	Yes					
Kosovo	PBA	Yes	Yes					Yes
Lebanon	PBA	Yes	Yes					
Lithuania	CBA	Yes	Yes	Yes	Yes			
Macao (China)	CBA	Yes	Yes		Yes		Yes	Yes
Malaysia	CBA	Yes	Yes				Yes	
Malta	PBA	Yes	Yes					Yes
Moldova	PBA	Yes	Yes					
Montenegro	CBA	Yes	Yes					
Peru	CBA	Yes	Yes	Yes	Yes		Yes	
Qatar	CBA	Yes	Yes					
Romania	PBA	Yes	Yes					
Russia	CBA	Yes	Yes		Yes			
Singapore	CBA	Yes	Yes		Yes			
Chinese Taipei	CBA	Yes	Yes		Yes		Yes	
Thailand	CBA	Yes	Yes	Yes	Yes			
Trinidad and Tobago	PBA	Yes	Yes					
Tunisia	CBA	Yes	Yes					
United Arab Emirates	CBA	Yes	Yes				Yes	
Uruguay	CBA	Yes	Yes		Yes			
Viet Nam	PBA	Yes	Yes					

\* Note by Turkey: The information in this document with reference to « Cyprus » relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the “Cyprus issue”.

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

\*\* B-S-J-G (China) represents the four PISA participating Chinese provinces: Beijing-Shanghai, Jiangsu, Guangdong.

Note: CBA = Computer-Based Assessment, PBA = Paper-Based Assessment. UH = “Une heure” shortened questionnaire version.

The context questionnaires contribute to integral aspects of the analytical power of PISA as well as to its capacity for innovation. Therefore, the questionnaire design must meet high methodological standards, allowing for the collection of data that leads to reliable, precise and unbiased estimations of population parameters for each participating country. In addition, the design also has to ensure that important policy issues and research questions can be addressed in later analysis and reporting based on PISA data. Both the psychometric quality of the variables and indicators and the analytical power of the study have to be taken into account when proposing and evaluating a questionnaire design. This is usually done by pre-testing all questionnaire content in the field trial one year prior to the main survey assessment. Accordingly, more material is tested in the field trial than will be implemented later on in the main survey. Results are then discussed with the PISA expert groups and material for the main survey is selected.

For PISA 2015, different assessment designs were implemented depending on whether a country used paper or computer-based tests. Only countries implementing the computer-based questionnaires were assessing the newly developed science material for PISA 2015. Countries using the paper-based assessment were mainly implementing trend material (i.e., material that was already used in previous cycles).

In addition, the field trial and the main study questionnaire designs greatly differ in many respects. The goal of the field trial is to evaluate the quality of the context questionnaires used in previous cycles as well as the quality of new items developed for PISA 2015. Moreover, processes and implementation are tested for all countries, including those that are new to PISA.



In the following sections, the differences between the field trial and the main survey design for both paper and computer-based assessments are explained in more detail.

## Field trial questionnaire design

### Computer-based design

For the Student Questionnaire, four parallel booklets were implemented. For the School Questionnaire, as well as for the optional Parent and Teacher Questionnaires, more material than could be used in the main survey was administered, leading to a slightly longer time to complete the whole questionnaire in the field trial than was planned for the main survey.

Each Student Questionnaire included a set of core items (i.e., StQ-FT Core Items) and one of four rotated blocks (i.e., StQ-FT-A, StQ-FT-B, StQ-FT-C or StQ-FT-D). The set of core items included a minimal set of student background variables – around five minutes in length – that were administered to all students. The four rotated blocks consisted of 25-minutes of non-overlapping content. As shown in Figure 17.2, these four blocks were randomly assigned to students. The optional questionnaires for students, Educational Career and ICT Familiarity questionnaires (ECQ and ICTQ) were administered following the Student Questionnaire and were available only as computer-based instruments.

The computer-based School Questionnaire in the field trial included trend and new material covering approximatively 60 minutes.

The optional computer-based Teacher Questionnaire covered approx. 45-minutes. It included a set of core questions (10 minutes assessment time) followed by two non-overlapping modules of 35 minutes each (TCQ-FTScience and TCQ-FTGeneral). The Teacher Questionnaire was administered to at most 10 science teachers and 15 teachers of other subjects in each school (For additional information about the sampling of teachers, please refer to Chapter 4).

### Paper-based design

Countries that chose the paper-based mode of delivery administered the paper-based Student Questionnaire. Students in these countries received both the tests and the questionnaires in paper-based forms. The paper-based Student Questionnaire took up to 30 minutes of assessment time and included mostly trend items, as well as some additional newly developed items.

The paper-based School Questionnaire included mostly trend items from previous cycles and is designed to be answered in approximately 60 minutes.

The optional Parent Questionnaire (PAQ) was administered on paper only, thus countries testing on paper as well as those testing on computer were able to implement this option. The PAQ included trend items as well as newly developed content and covered an assessment time of approximately 30 minutes.

The field trial questionnaire designs for the Student Questionnaire and the Teacher Questionnaire are shown in Figure 17.2 below.

■ Figure 17.2 ■

### Field trial computer-based design for Student (StQ) and Teacher Questionnaires (TCQ)

Student Questionnaire			
StQ-FT Core Items (5 min): gender, age, grade, educational program, parental occupation, parental education, immigration background			
Within-school random assignment to one out of four non-overlapping blocks (25 min each)			
StQ-FT-A	StQ-FT-B	StQ-FT-C	StQ-FT-D
(Optional) Educational Career Questionnaire (10 min)			
(Optional) ICT Familiarity Questionnaire (10 min)			
Within-school random assignment to one out of two non-overlapping blocks			
ICT-FT-A			ICT-FT-B

### Optional: Teacher Questionnaire

TCQ-FT-Core: Teacher background, school climate (10 min)

TCQ-FT-S (35 min)

Administered to the sample of science teachers

TCQ-FT-G (35 min)

Administered to the sample of non-science teachers

## Main survey questionnaire design

The questionnaire designs for the field trial and the main survey were different. The main survey Student Questionnaire consisted of only one booklet and the assessment time was again limited to a maximum of 30 minutes. The School Questionnaire content was also reduced to an assessment time of approximately 45 minutes. The questionnaires in total still covered all policy modules proposed in the questionnaire framework (see Chapter 3). The two optional questionnaires for students – Educational Career and ICT Familiarity – were kept at 10 minutes in length each.

The mode of assessment did not change from the field trial to the main survey, i.e. countries that implemented the assessment on computer also administered the computer-based questionnaire, while paper-based testing countries administered a limited set of mainly trend questions for students and schools. The Parent Questionnaire again was administered on paper only, while the Teacher Questionnaire and the optional ICT and Educational Career Questionnaires were available only on computer.

The main survey questionnaire designs for the computer-based instruments are shown in Figure 17.3 below.

■ Figure 17.3 ■

### Main survey computer-based design for Student (StQ) and Teacher Questionnaires (TCQ)

**Student Questionnaire**  
(n = 6300 in CBA Design per country)  
(approximately 30 minutes)

Optional: Educational Career Questionnaire (ECQ) (10 min)

Optional: ICT Familiarity Questionnaire (ICTQ) (10 min)

**School Questionnaire (ScQ)**  
(n = 150 per country)  
(45 min)

**Optional: Teacher Questionnaire (TCQ)**  
(up to 10 science teachers and 15 non-science teachers per school)  
(30 min)

TCQ-MS-Core: Teacher background and education (5 min.)

TCQ-MS-S (25 min)

Administered to the sample  
of science teachers

No overlap with TCQ-MS-G

TCQ-MS-G (25 min)

Administered to the sample  
of non-science teachers

No overlap with TCQ-MS-S

As the majority of countries decided to implement the computer-based assessment for this cycle, the next paragraphs describe the computer-based questionnaires in more detail. The description of steps 2 to 8 of the questionnaire life cycle focuses on the questionnaire platform and the associated functionalities.

## STEP 2: MASTER QUESTIONNAIRES AUTHORIZING

The implementation of the cycle described in the previous section is supported by a set of tools, integrated in two major subsystems, in the PISA platform.

The first subsystem is the PISA portal, step 1 (master questionnaires design) and related activities (e.g. general information sharing, files sharing and global tracking of issues using the PISA platform).



The second subsystem is the PISA questionnaire platform, a comprehensive toolbox that focuses on: production (i.e. their definition, authoring, testing, adaptation, and validation) of questionnaires (master and national questionnaires), delivery of these questionnaires to respondents, and the management of all administrative aspects related to them.

Consequently, the questionnaire platform is designed to reflect these goals. When users log in to the platform, they are taken to a home page as shown in Figure 17.4, providing them access to the platform's features.

■ Figure 17.4 ■  
**Questionnaire platform home page**

## Questionnaire authoring tool

### **Main view for questionnaire editing**

Users working on questionnaires first see the questionnaire authoring tool (QAT) editor when connecting to the questionnaire platform. The tool is used to author the computer-based questionnaires of PISA 2015. It is an online editor that allows a user to add, suppress or edit a question. When users open the QAT editor, they are presented with a view on the structure of an entire questionnaire, which is not a *what you see is what you get* (WYSIWYG) view of what participants eventually see. Figures 17.5 and 17.6 show the main view of this editor for a National Project Manager (NPM).

■ Figure 17.5 ■

**QAT main view (with a specific question SC002 as an example)**

**English (Australia) For School\_Questionnaire > Production**

Check variables Export PDF Cancel last changes Save Home Log out

Navigation Items Errors

Screen 1 of 37 ID: SC0info

Screen 2 of 37 ID: SC1info

Screen 3 of 37 ID: SC001

Screen 4 of 37 ID: **SC002** Template: List of text inputs (+ pie chart)

Screen Options: Hidden:  Invert Columns and Rows:  Pie Chart:  Not included in List of Items:

Question: As of 1 April, 2015, what was the total school enrollment (number of students)?

Description:

Help:

Instruction: (Please enter a number for each response. Enter "0" (zero) if there are none.)

Responses:

Code	Response	Input size
Q01TA01	Number of boys	Medium
Q02TA01	Number of girls	Medium

Format: Integer  
Min value: 0  
Max value: 10000

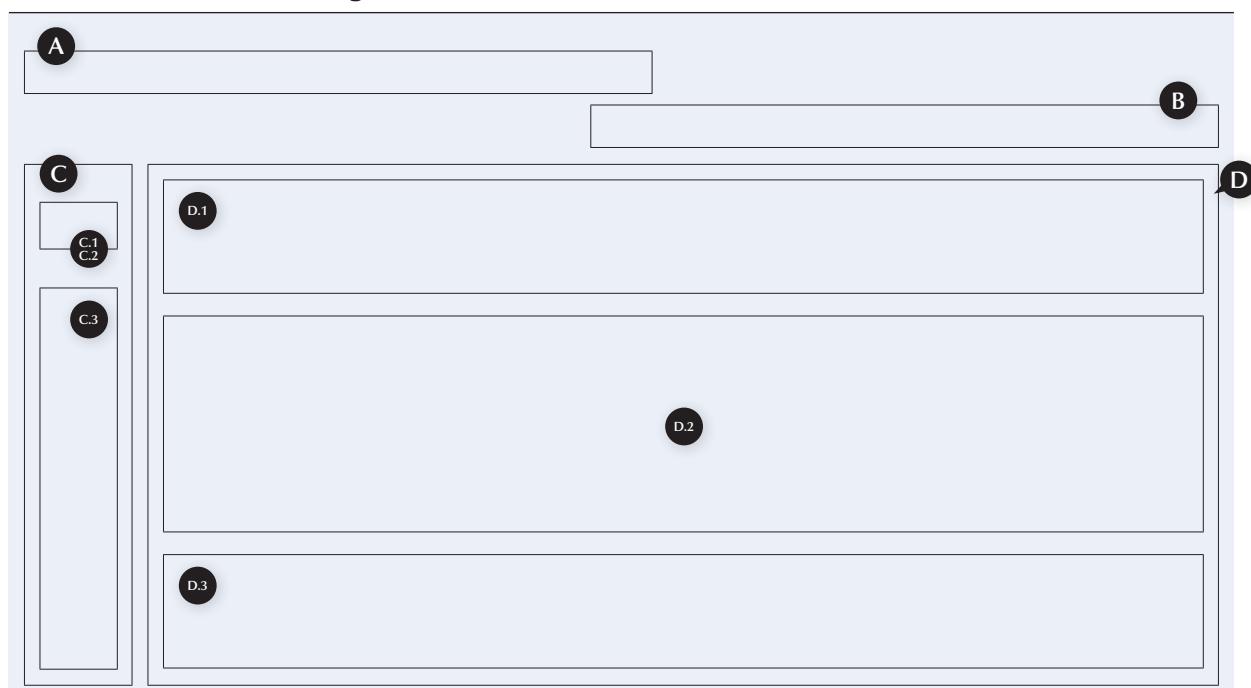
Footer:

Screen 5 of 37 ID: SC003

Screen 6 of 37 ID: SC004

Copyright © 2013 - CRP Henri Tudor - All rights reserved

■ Figure 17.6 ■

**Organisation of the main view of the QAT editor**



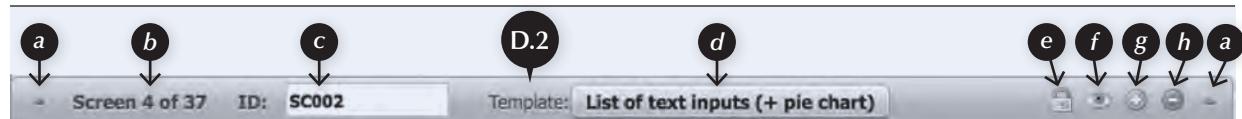
The organisation of the main view presented in Figure 17.6 is the following:

- A. **The questionnaire title** concatenates the questionnaire label (country, language and type of questionnaire) and the questionnaire mode (i.e. the modes of the QAT are critical as they define the rights of a current user. According to the mode, the access for modifying questionnaires in the QAT editor is locked or unlocked, allowing each user to work independently).
- B. **The questionnaire toolbar** provides the following action buttons:
  - *Check variables* checks throughout the questionnaire if an identifier is used by more than one variable (this check is also automatically performed when the *Save* action is triggered).
  - *Export PDF* generates a PDF version of the questionnaire.
  - *Cancel last changes* reloads the previously saved version of the questionnaire.
  - *Save* saves the questionnaire to the database. When used, this action triggers two kinds of checks: one to check if all rules are correctly formatted (no missing variables, no syntax error) and the other one to check if each variable has a unique identifier. If one test fails, the questionnaire is saved but the user will be unable to execute it.
  - *Home* redirects to the questionnaire platform home page.
  - *Log out* disconnects the user from the platform.
- C. **The navigation menu** is a panel offering two viewing options:
  - a list of the question items [C.1] or
  - a list of the unresolved errors (e.g. problematic rules) [C.2]
  - and quick access to the related question [C.3].
- D. **The QAT editor** displays the list of all questions (called “screens”) and rules (called “rules headers”) available for a questionnaire. When clicked, each part toggles between an expanded (D.2) or a collapsed view (D.1 and D.3).

#### **Questions expanded view and questions preview**

Figure 17.7 shows the features available at the top of the expanded view.

■ Figure 17.7 ■  
**The expended view information**



- a. *Show/hide screen* button toggles between a collapsed or expanded view of the question (screen or rules header).
- b. Screen NUM of TOTAL (where NUM is the rank of the screen in the sequence of screens of the loaded questionnaire in the QAT editor and TOTAL is the total number of screens existing for the edited questionnaire) or *rules header*.
- c. The *ID* field displays the technical identifier of the screen or rules header.
- d. The *template selector* displays the name of the template used for editing the question (see part below about the question templates).
- e. The *lock* button, not made available to National Project Managers (NPMs), gives them the right to edit or not.
- f. The *preview* icon opens a preview of the item.
- g. The *add screen* icon inserts a new question or rule in the edited questionnaire.
- h. The *delete screen* button removes (after confirmation) the question or rule from the edited questionnaire.



The questionnaire platform offers two preview options for reviewing and checking the quality of the masters encoded in English (Figure 17.8).

■ Figure 17.8 ■  
**Preview of a question with the QAT editor**

The first option is a question preview panel triggered within the QAT editor, via the preview icon available in the expanded view of each question. In this preview mode, the identifiers of response fields are visible to facilitate the questionnaire authoring.

The second option is a full questionnaire preview accessible via the *runtime* menu entry of the questionnaire platform home page. This option lets users navigate through a questionnaire in a test environment and offers the same conditions as those met by the “real respondents” when the questionnaire goes to the field.

### **Question templates**

Inside the expanded view, the user can edit the different parts of a question using the QAT editor: the question text, the description, the instruction, the help and the different answer categories.

The QAT editor is a template-based questionnaire authoring system that supports, amongst other features, the creation of multilingual contents (including left-to-right and right-to-left written texts, extended character sets for Arabic, Chinese, Hebrew, Japanese, Korean, Russian, Thai, etc.), the design of the rules-based routings driving the questionnaire flow, and the enforcement of the quality of the answers via validation rules and constraints.

The question types – or the templates – available in the QAT editor are:

- *exclusive choice*
- *multiple choice*
- *list of text inputs (+ pie chart)*
- *list of exclusive choice (table)*
- *list of multiple choice (table)*
- *multiple list of text inputs (table)*
- *simple list of text inputs with check-in option*



- scale question type (also called slider)
- free text input
- forced choice
- drop down list
- drop down (table)
- information.

Additionally, there are two templates for defining rules that are used within the questionnaire:

- consistency check rule
- routing rule.

A short description of each template is given below, with examples in Figures 17.9 through 17.20.

■ Figure 17.9 ■  
**Information template**

### Information

***On this screen, You may insert general information, useful for introductory items.***

***There is no interaction inside such items.***

***Formatting using :***

- bold
- italic

The *information* template is used to insert an introduction, a transition or a closing page into the questionnaire.

The author can use this template to present the questionnaire (e.g. its goals, structure, general recommendations and other instructions...), to introduce a new section of questions and to thank the respondent at the end of the questionnaire.

■ Figure 17.10 ■  
**Exclusive choice template**  
(technical name *simpleMultipleChoiceRadioButton*)

### Exclusive Choice

**How many computers are connected to...**

*If there is no computer in your sch...*

None

From 1-5

The *exclusive choice* template presents a question to the respondent as well as a set of mutually exclusive responses.

Each response option receives an identifier. The data saved for this template includes a value of either 0 or 1, for each response option. At most, only one of these values will be 1.



The presentation of this item type to the respondents uses a single set of standard radio buttons. Choosing one of the options will remove any previous choices.

■ Figure 17.11 ■

### **Multiple choice template**

(technical name *simpleMultipleChoiceCheckbox*)

The *multiple choice* template presents a question to the respondent as well as a set of non-exclusive responses.

Each response option receives an identifier. The data saved for this template includes a value, either 0 or 1, for each response option.

The presentation of this template uses standard checkboxes. The checkboxes are selected (with a checkmark or X) when a user clicks on them, and unselected if clicked a second time.

■ Figure 17.12 ■

### **List of exclusive choice (table layout) template**

(technical name *complexMultipleChoiceRadioButton*)

	never	Once a year	Once a month	Everyday
Verbal aggression	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

This template presents the user with a set of exclusive choice questions on a single screen in a tabular format. In the default format, each row of the table is a separate response, and the columns are a set of choices for each response. In addition, the QAT editor allows the author to invert the table, so that responses are in the columns and the choices are in the rows.

Typically, this template presents a single question text (e.g. To what extent do you agree with the following statements?). The choices in the columns indicate a range for the responses (e.g., from strongly agree to strongly disagree), and the responses gathered in each row indicate one specific aspect (e.g. a statement) that should be evaluated by the respondent.



In the default case, where responses are in rows, each row will be a set of radio buttons. Clicking on one of the radio buttons will clear any previous choices in that row. A data value is collected for each radio button on the screen. So, if there are 4 rows and five columns, a total of 20 data values will be collected.

■ Figure 17.13 ■

### List of multiple choice (table layout) template

(technical name *complexMultipleChoiceCheckbox*)

**List of multiples choice (table)**

**How decisions are taken with respec...**

	Parents	Teachers	Principals	IT staff	Students
<b>Who is involved in</b>	<input type="checkbox"/>				

This template presents the respondent with one or more non-exclusive choice questions on a single screen in a tabular format. It is similar to the previous template; however, it uses checkboxes so that more than one choice can be selected for each row (or column if the presentation is inverted).

■ Figure 17.14 ■

### List of text inputs (+ pie chart) template

(technical name *simpleFieldsList*)

**List of text inputs (+ pie chart)**

**What is the funding Scheme of you S...**

<p>Public funding</p>	<input type="text"/>
<p>Private funding</p>	<input type="text"/>

This template is used for collecting short, open ended response data. The template presents the respondent with one or more areas to type a response, each with a label indicating the information to be entered.

The responses can be unfiltered text, or they can be limited to numeric values. Constraints can be placed on the values entered in each case. If unfiltered text is allowed, the response can be limited to a minimum and/or maximum length of text. If the response is numeric, a minimum and/or maximum numeric value can be specified. If respondents give a response outside the permitted ranges, an error message is displayed.

An optional feature of this template is the ability to include a pie chart as part of the presentation. This pie chart is constructed dynamically as the respondent enters values into the response areas. Each response area corresponds to a section of the pie chart. The responses must be numeric, and if the sum is greater than 100, an error is shown.

■ Figure 17.15 ■

**Multiple list of text inputs (table layout)**  
 (technical name *complexFieldsList*)

	Classic	Science
Cycle 1 students		

This template, like the previous one, is used for collecting short, open ended response data. However, in this case more than one response can be collected for each area of interest. The response areas are presented as a table. Similar to the previous template, the response values can be either text or numeric, and can be limited in their range.

■ Figure 17.16 ■

**Scale question type template**  
 (technical name *slider*)

The *slider* is one of the innovative interaction models used in the PISA2015 platform. It facilitates the work of the questionnaire author who needs to collect a relative value within a given range. The respondent moves an indicator along a scale line to indicate where in the range their answer should be.

The template allows the author to include one or more slider responses on a screen. Each slider has upper and lower limits which are integer numbers. The author can include labels for the left and right ends of the scale. Also, the step value for the slider can be set. By default, the step is 1, so each integer value in the range can be selected. But this step can be changed to, for instance, 10, which would only allow answers that are incremented by 10.



■ Figure 17.17 ■

**Free text input template**  
(technical name *textfield*)

This template supports an open ended text response. The respondent is presented with a large box in which he can enter a long text and is able to include line breaks to provide multiple paragraphs in the response.

This template was not used in any of the PISA 2015 master questionnaires, but it was used by some countries for their national extensions.

■ Figure 17.18 ■

**Forced choice template**  
(technical name *multipleItems*)

a) When I study a\_01 for a mathematics test, I try to

a\_02 for a mathematics test, I try to

a\_03 for a mathematics test, I learn

This template is similar to the *exclusive choice* template. It presents the user with one or more questions with multiple answer options. Each question can have one and only one option selected. The primary difference between the two templates is in how they are formatted on the screen. In the case of *forced choice*, a descriptive text is presented at the top of the screen, then for each question the choices are displayed in a row horizontally. This template was mainly used for trend questions of previous cycles.

■ Figure 17.19 ■

**Drop down**(technical name *simpleDropDown*)

**Drop Down**

**How many computers are connected...**

When I study for a mathematics test...	SC02_01	
	Un	+
When I study for a	SC02_02	

This template presents the respondent with one or more drop down menus from which to select their response. Each menu can have a textual label to present a question or to label the contents of the menu.

The contents of the menu are defined using some lists. The menus can share the same list of response values, or each can have a unique list. For instance, a question could ask for the date of birth, with three different drop down menus for the day, month and year parts of the date.

■ Figure 17.20 ■

**Drop down (table layout) template**(technical name *complexDropDown*)

**Drop Down (Table)**

**How many computers are connected...**

	Teacher	Student
When I	SC02_01_01	SC02_01_02
study for a mathematics	Un	+

Like the *drop down* template, this template presents the respondent with one or more drop down menus for providing a response. In this template, the menus are organised into a table presentation.

Like the previous template, the drop down menu contents are defined in one or more lists. In the standard layout, each menu in a row will contain the same list of response values. However, like the other table based templates, it is possible for the author to invert the rows and columns so that columns contain the same menu values.

**Consistency check rule**

The consistency check rule template supports a rule-based approach for validating the response provided by a user. The author provides a Boolean condition (i.e. “true” or “false”, intended to represent the truth values of logic) that checks the values of some response variables from different questions the respondent has answered. If the condition evaluates to TRUE, a message is displayed to the user.



The template for defining the consistency check rule appears as follows:

■ Figure 17.21 ■

### Consistency check rule template

The rule is evaluated when the respondent navigates away from the current question (e.g. by clicking *next* or *log out*). When the condition is true, a message is shown like the one below:

■ Figure 17.22 ■

### Consistency check message



The respondent can click on “Ok” and go back to the current question to change his or her response. If the respondent clicks the “Skip the check” button, the navigation proceeds as normal.

### **Routing rule**

The routing rule allows the author to use branching within a questionnaire. Routing rules appear in between questions in the questionnaire. They are executed after the completion of the question before the rule.

The routing rules are based on Boolean conditions, similar to the consistency checks. The rules are defined using an IF-THEN-ELSE logic. If the condition evaluates to TRUE, the THEN part is executed, otherwise the ELSE part is executed. The THEN and ELSE parts can be either another IF-THEN-ELSE rule (allowing nested logic to be defined) or GOTO commands, directing the questionnaire runtime to branch to a specific question in the questionnaire.

The routing rules are typically used for skipping questions that do not make sense given a specific initial response from the respondent. A simple case is an exclusive choice question, where the last response option is “other”. If the respondents select this option, they should be shown a question asking for more information about their answer, e.g., an open ended response where they can type in their answer. Such a rule could be defined as follows:

■ Figure 17.23 ■

### Routing rule template

In this case, ST019 would be the initial question where a respondent has the choice to select “other”. ST021 would be the follow-up question asking for more information, and ST022 would be assigned if ST019 has not been answered with “other”.

#### **Concept of questions and answers identifier within the QAT**

An identifier (or ID) is a tag attached to an object. The ID allows the object to be referenced and to be retrieved and used in a precise perimeter of action, or scope. A relation between a tag and the object that it references must be unequivocal. Consequently, the label given to an identifier must be unambiguous and unique within the perimeter where the referenced object can be used.

In the QAT editor, the types of objects receiving an ID are the various questions, including the rules, and all elements designed to receive and store the data provided by the respondents (i.e. answers).

The IDs are one of the key parts for the computer-based questionnaires and are the basis for the data analysis. A question (or part of question) with an unexpected or inappropriate ID is unusable and can eventually not be analysed. Checking the consistency of IDs is one of the main important tasks done by contractors when reviewing a computer-based questionnaire.

### **STEP 3: CREATION OF NATIONAL QUESTIONNAIRES**

As soon as the master questionnaires are authored and checked, they are duplicated for each country and national language version, so everybody starts with the same basis. These questionnaires become the first version of the national questionnaires. The copy operation is performed by the technical team of the PISA questionnaire platform using several system scripts. These national questionnaires are then put into a mode that allows the national centre to adapt and translate the content, as described in step 4 of the questionnaire life cycle.

For each national questionnaire, the users continue to have access to the corresponding version of the master questionnaire in a “read-only” mode via the “Open Master” menu entry of the questionnaire software home page. To facilitate the work (i.e. reference, comparison, etc.) of the user, this “read-only” master questionnaire is displayed in a new tab-page or a new instance of the web browser.

### **STEP 4: NATIONAL QUESTIONNAIRE ADAPTION AND TRANSLATION**

The main work performed at this step is done by the national centre within a country. Once the national questionnaires are ready, the national centre has edit access to it in order to integrate their agreed adaptations and reconciled translations. Contrary to the cognitive assessments, the use of professional text translation formats (e.g. XLIFF formats) is not used for the questionnaires as the very last version of the translated questionnaires is directly integrated in the QAT editor. Like for authoring the master questionnaires, the national centre has access to the same functionalities in the QAT editor, such as adding new national questions and adapting existing questions, as well as the functionalities for previewing the questions. A functionality called “Copy item between questionnaires” can also be used in order to copy some questions from one questionnaire to another one. Thus, the same translated question only needs to be integrated once in the QAT editor.

When opening the questionnaire, the national centre can see the master questionnaire texts in English as well as some of the national questions (or parts of questions) already translated and locked. These locked questions are called “trend questions” and represent the questions used in previous PISA cycles. Maintaining the quality and integrity of the trend questions over time is important to be able to analyse the data across cycles. Thus, the verifiers take the paper version of the questions from the previous cycle and manually transfer it in the QAT editor before the national centre gets access to their national questionnaires. Then, using the lock buttons, the verifier locks the questions. These questions will appear in orange and indicate that they cannot be edited anymore. If the national centre wants to modify such an item, they must negotiate the adaptations or requested changes with the questionnaire content experts for these trend questions. If these changes are accepted, the verifier will make the changes for the national centre and again lock these questions afterwards. This translation and verification process is described in more detail in Chapter 5 of this report.

By default, the questionnaire software proposes a set of automatic formatting adapted for the PISA questionnaires such as questions displayed in bold, instructions in italic, etc. However, some of this automatic formatting might need to be adapted by the national centre according to their cultural specificities. For example, a standard font size fit to Latin-based character sets may be too small to display the intricacies of Chinese Kanji characters. Therefore, the questionnaire



software includes a function that allows the national centre to customise settings in their language. With this functionality, the user can specify the text reading direction, the font family, the font size, the text styles (bold, italic, underline, etc.), the line height and the text alignment. This configuration tool is accessible via the “Runtime Style Authoring Tool” menu entry of the questionnaire software home page.

For the table templates, the users are also able to adjust the column widths to optimise the display of each question. This feature can be useful for languages that have long words such as the German language. These adjustments are available when previewing the individual questions in the QAT editor, in a WYSIWYG mode. It is the only screen layout changes that are allowed for the computer-based questionnaires.

When all the translations have been inserted, the fonts are validated and the layout is checked, the national centre can test their questionnaires and validate their work. This part is completed via the “Runtime” menu entry of the questionnaire software home page.

### **STEP 5: NATIONAL QUESTIONNAIRES QUALITY CHECK**

The quality of the national questionnaires must be checked according to different views: the quality of the translations, the accuracy of the translation compared to the English master version, the respect of the agreed adaptations and the technical validity of the questionnaires.

At this step, most of the checks are done manually and each contractor gets access to the questionnaire platform in a read access mode.

The translation and adaptation discrepancies are documented in an Excel file which is delivered to the national centre for their review. The national centre is therefore able to accept or refuse these comments and can update their questionnaires accordingly. (See Chapter 5 for a more detailed description of the translation validation).

The technical team of the questionnaire platform is also involved in this step to manually check all the questionnaires according to several criteria: making sure a user is able to go through the questionnaire from the beginning until the end without a software error due to, for instance, errors in routing rules, check if the number of questions match the number of agreed questions, check if all questions and messages are translated, check if all the parts of the interface are translated and well integrated and check all IDs to make sure they are in agreement with the master. As explained in step 2 of the PISA questionnaire cycle, IDs are the key identification point for the data analysis and an error in this part might result in loss of data.

National centres are provided with testing scenarios for each questionnaire to validate the accuracy of their work. These scenarios describe different ways in which a respondent could answer a questionnaire following every possible routing. National centres are asked to test the questionnaires several times based on these scenarios. When national centres are done with their testing, they need to send their results to the technical team who will analyse all the files and make sure that no technical problems are detected as it is the last step before going to the field.

As all activities performed by the national centre are carefully saved, the technical team is able at any time to monitor the different activities and help in case of technical issues. The technical support provided required 24/7 availability due to the different time zones covered in PISA.

### **STEP 6: PREPARATION OF NATIONAL QUESTIONNAIRES FOR DELIVERY**

After the different checks and controls are performed, the translation and verification of the questionnaires are completed; all is ready to be delivered to the respondents. At this step, the QAT administrators and technical team make a number of checks and system setups using the questionnaire platform’s administrative interface shown in Figure 17.24.

■ Figure 17.24 ■

**Questionnaire platform – administrative view**

The screenshot shows a web-based administrative interface for the PISA 2015 questionnaire platform. At the top, there is a header bar with '[User]' on the left and '[Log out]' on the right. Below the header, a large box displays the message 'Welcome to your PISA 2015's administration page'. Underneath this, a section titled 'Please make your choice' lists various administrative tasks. Each task is preceded by a bullet point and a bold title, with sub-options listed below them.

- **Users Monitoring Overview**
  - Real Time Users Monitoring
  - History Users Monitoring
- **Questionnaires Versions Management**
  - Questionnaires Backup
  - Promote a Questionnaire
- **Profile management**
  - Manage Designs
- **Countries Manuals Management**
  - Manage School Questionnaire Manual
- **Sampling management**
  - Manage Users School Questionnaires
  - Manage Users Teacher Questionnaires
  - Import Sampling Files
- **Countries Monitoring Management**
  - Manage monitoring access
  - Overview Countries Monitorings
- **USB Stick Management**
  - USB Key Generator
- **Clusters Management**
  - Country export
  - Cluster History
- **Administration Tools Management**
  - Langs Settings
  - Display roles
  - Add "QAT Author" to all users
  - Check Questionnaires
  - Show All Empty Path
  - Show Translation State

There are two modes of delivery used for the questionnaires. For student questionnaires, including the optional ICT and EC questionnaires, the questionnaires run in an offline, standalone mode as part of the PISA student delivery system (SDS). The School and Teacher Questionnaires are delivered online over the web to respondents around the world. Both modes share a common code base and database structure, but the preparation for delivery follows different procedures.

For the student questionnaires, the preparation step primarily involves exporting the completed national questionnaires for each country, as well as the questionnaire software and user interface translations, in a form that can be used on USB drives for delivery. Unnecessary components, such as the QAT editor, are removed from the questionnaire platform, and a database image with the national questionnaires is created. These exported files are directly integrated into the PISA SDS software for a country, and then tested and validated. See Chapter 18 for more information about the student delivery system (SDS).

The online School and Teacher Questionnaires require more steps to prepare for the field. A key step is to import the sampling information into the questionnaire platform so that the selected schools and teachers will be known to the

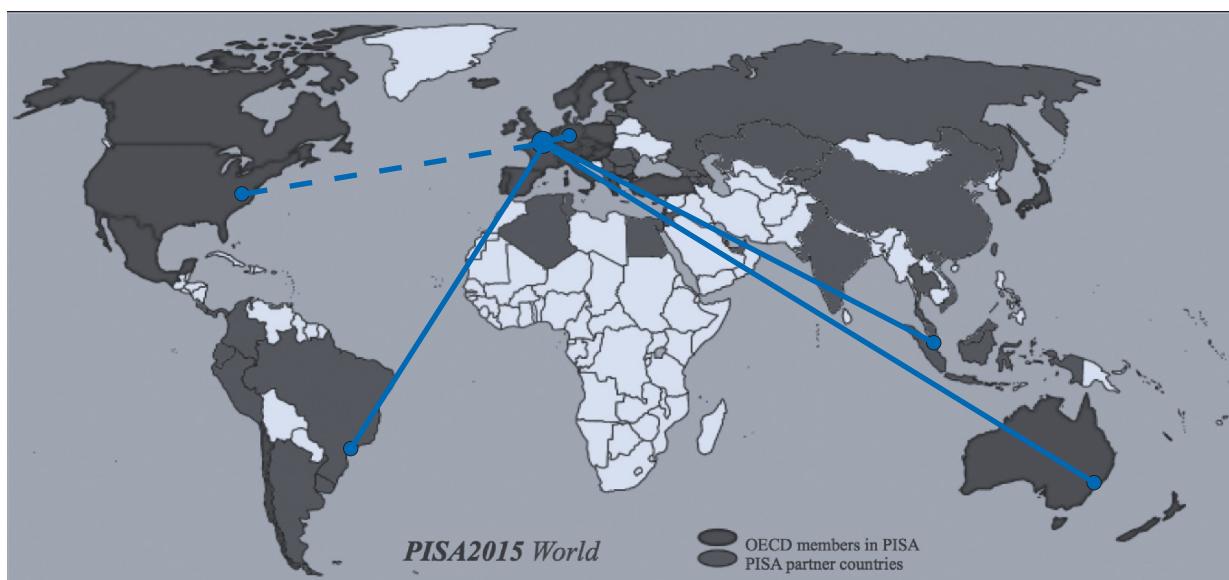


system and can be identified when they connect to complete the questionnaires. To do this, the “Sampling Task 5” (for the field trial) or “Sampling Task 11” (for the main study) output files are taken from KeyQuest, the system used for sampling within countries (see Chapter 4 for details about the sampling). These files contain the list of schools selected from the sampling process, using anonymised ID codes. They also contain information that describes the range of IDs that will be assigned to teachers if the country participates in the optional Teacher Questionnaire. The files are imported into the questionnaire platform, which creates logins and passwords for each sampled school and teacher. These logins and passwords are then sent to the national centre, which distributes them to the selected schools and teachers.

The countries participating in the online questionnaires in PISA 2015 were spread around the world. For the field trial, a single server, located in Luxembourg, was used for data collection. For the main study, in an effort to improve the performance for the end users, the questionnaire platform was distributed to servers around the world (shown in Figure 17.25). This helped to reduce the network latency that users experienced, and improved perceived performance. In addition to the primary server in Luxembourg, two server installations were added in Frankfurt, Germany, and one server was added in each of the following locations: Singapore, Sydney, Australia and Sao Paolo, Brazil.

■ Figure 17.25 ■

#### Distribution of the PISA 2015 servers



Countries were assigned (in a transparent way) to one of these server locations. Respondents were given a URL which connected them to the primary server in Luxembourg. Based on the ID used to login to this server, the system could determine which country the user came from, and which server they should be assigned to. The respondent was then automatically redirected to this server, where they would take the questionnaire.

In addition, one country, the United States, delivered the online questionnaires from their own national server. This server was completely standalone, so respondents connected to it directly, not through the central PISA server in Luxembourg.

Each location in this network of servers was composed of a tandem of servers mounted in a Master-Slave mode with a failover database mechanics guaranteeing the security of the data in case of a Denial of Service attack or because of a system or software failure. The slave server was also used for generation of results files (see step 8) for performance reasons.

#### STEP 7: DATA COLLECTION AND QUALITY MONITORING

In step 7 of the questionnaire life cycle, results are collected from students, school principals and teachers. The respondents proceed through the questionnaire, seeing the same rendering and behaviours as the QAT authors have when previewing the questionnaires in the questionnaire platform. For the students, this is done as part of the PISA student delivery system, typically running from USB drives on school computers. The questionnaire software runs offline, in a standalone mode on the school computer, and all results are saved back to the USB drive. The students do not need

to login to start the questionnaires. Identification and authorisation of the students is performed by the student delivery system.

For the online questionnaires for school principals and teachers, delivery is performed over the Internet. This requires the principals and teachers to identify themselves prior to beginning the questionnaire. Respondents are assigned login IDs and passwords as part of the sampling process in step 6. When they first connect to the questionnaire platform, they must enter this ID and password. The questionnaire software will select the appropriate questionnaire based on this ID. In some countries, users must select which language they would like to use when completing the questionnaire.

As respondents complete the questionnaires, data is collected by the questionnaire platform. The original data saved is the response to each question. This data depends on the template used for each question. For questions that use radio buttons or checkboxes, a data value is saved for each of these controls on the screen. The value will be zero or one depending on whether the control has been selected. For sliders, dropdown menus and textual responses, the value selected or entered is saved. If no response is selected or entered, a value of “null” is saved.

Along with the response data, additional data is saved for each respondent. The final valid path taken by the respondent in the questionnaire is saved. This allows one to determine which questions are valid and were presented to the respondent based on the routings that were taken. Also, a log of actions by the respondent and the questionnaire system is saved. This log includes events such as those shown in Figure 17.26.

■ Figure 17.26 ■

#### Logged events

SESSION_START	The user starts or resumes a questionnaire
ITEM_START	The user starts an item
HELP	The user clicks the Help button
RESET	The user clicks the Reset button to clear previously entered answers
LIST_OF_ITEMS	The user clicks the List of Items button to see the questions that have already been visited in the questionnaire
SELECTED_JUMP	The user clicks on one of the questions in the List of Items to jump to that item
SELECTED_FORWARD	The user clicks the Next button to move forward in the questionnaire
SELECTED_BACK	The user clicks the Back button
SELECTED_LOG_OUT	The user clicks the Logout button to leave the questionnaire
MOVE_FORWARD	The system moves forward to the next question
MOVE_BACK	The system moves back to the previous question
MOVE_JUMP	The system jumps to a new question
LOG_OUT	The system logs off the user
ANSWER_SELECTION	An answer is selected or entered
RANGE_CHECK	The answer entered triggered a range check
CONSISTENCY	A consistency error message is displayed
CONSISTENCY_CANCEL	The move action is cancelled due to the consistency error
CONSISTENCY_SKIP	The consistency error is skipped and the move action proceeds

During this phase, for online questionnaires, the National Project Managers and administrators of the questionnaire platform can monitor the activity of the questionnaire respondents. The monitoring shows which respondents have connected to the questionnaire platform and how far they have progressed through the questionnaire. The platform also supports generating a PDF file for a respondent, showing the questionnaire including all the responses that have been saved. The overall status information can be exported to a spreadsheet for further sorting and filtering.

In the main study, the sampling process selects schools that are chosen to participate in the PISA survey, along with replacement schools if the originally sampled schools refuse or are unable to participate. Through the monitoring tools available in the questionnaire platform, the NPMs are able to activate and disable these schools to control access based on their status. Additionally, some countries used this feature to disable schools after they have completed their questionnaires.

The administrators of the questionnaire platform have additional tools available for monitoring the progress of the respondents. These include a view of all currently connected users, as well as a history of the logins, both successful and unsuccessful. These reports are important in supporting users who report problems and also for monitoring performance issues on the servers. Additionally, the questionnaire platform saves many different logs, which the administrators use for detecting problems and troubleshooting them. All these servers are monitored and must be up 24/7.



## STEP 8: COMPLETION OF DATA COLLECTION

Following a negotiated agenda depending on a country's testing date, the access to the online questionnaires is closed. The production phase of the national questionnaires is then ended. This fixed end date allows the final export of results data for inclusion in the PISA analysis. After the access is closed, respondents who attempt to login receive a message indicating that the questionnaires are currently not available and asking them to contact their national centre for further information.

Each country's result data is exported on a weekly basis. Due to the large volume of data, the data generation is performed only once a week to reduce the load on the system. The national centres can download the latest results in a single compressed file, which is imported directly into the Data Management Expert system.

The access to the servers and the questionnaire software is available several weeks after the end of the data collection to allow some time for the NPMs to retrieve the data but also ask questions in case of problems with the data collected.

## DEVELOPMENT PROCESS OVERVIEW AND TECHNICAL INFRASTRUCTURE

This section describes the technical aspects of the software and hardware used to support the computer-based PISA 2015 Questionnaires. The PISA Questionnaire platform is a complex and relatively large software system. The development followed standard software development processes. A modified agile process (see [https://en.wikipedia.org/wiki/Agile\\_software\\_development](https://en.wikipedia.org/wiki/Agile_software_development)) was used, implementing multiple releases in the course of developing the platform. An open source project management platform (Redmine, <http://www.redmine.org>) was used to track and document the work.

The PISA questionnaire software was written primarily in PHP on the server side and JavaScript within the web browser. The Apache web server was used for delivery of web content, and data was saved using the MySQL database system. The questionnaire content was structured using custom XML markup.

The online questionnaire servers were Linux based, using Ubuntu 12.04 LTS. The student questionnaires were delivered as part of the PISA student delivery system, which was based on XAMPP. For the main study, multiple servers were deployed using the Amazon Web Services EC2 system.

The methods for software testing evolved as the project progressed. Aspects of unit testing (using the Jenkins system, <https://wiki.jenkins-ci.org>) were implemented, but the fundamental testing was functional and integration testing performed by developers and project managers. A system of automated functional testing using a farm of more than 40 computers running various web browsers and OS's was also deployed. Finally, load testing of the online questionnaires was implemented using the JMeter system (<http://jmeter.apache.org/>).

## CONCLUSION

As we saw with the description of the steps of the PISA 2015 questionnaire life cycle, having computer-based questionnaires provided several advantages: flexibility to accommodate language constraints, easy monitoring and check of the work done by the users, a more efficient and reliable data collection process, and collected data available quickly and cleaner for the final analysis. While the advantages of switching to the computer-based questionnaires are important and are a significant motivator to make the PISA cycles more innovative, it is important to also recognise the challenges that countries faced with online delivery of questionnaires. Having access to the Internet and a web browser, which is a basic part of a modern society in 2015, is a necessary but not sufficient requirement to be a part of a computer-based study. The major challenge with delivering online questionnaires taken by thousands of people around the world who can be anywhere (at home, at work, in a cyber cafe...) is that the environment is not controlled at all and all problems cannot be anticipated. For example, what kind of definition should we give for "reliable Internet connection"? Just having a look on Wikipedia for the average connection speed in different countries shows the huge range of Internet infrastructure in the PISA countries.<sup>1</sup>

For PISA 2015, several national centres had to deal with "technical issues" such as users who were unable to have access to a computer with the minimum web browser version supported by the study, or users who refused to continue to answer the questionnaire due to a very slow Internet connection. Other technical challenges encountered by users included network filters in schools that interfered with access to the questionnaires and web browser extensions that interfered with the web pages that implement the questionnaires. Finally, some users could not meet the relatively modest requirements for browser versions. In general, these problems are common to any large scale web application.



The consequences of these problems are a reduction in response rates and lost data as users are reluctant to take part in a questionnaire which is seen as difficult to answer. Some national centres had to send paper versions of the questionnaires to principals and teachers to increase response rates. This increased the workload of the national centre and reduced the value of an online survey.

While most countries already realise the benefits of transitioning from paper-based to computer-based questionnaires, paper-based questionnaires still have their place in PISA, and will for the foreseeable future as some countries still need further development of the infrastructure required to support online questionnaires.

### **Note**

1. [https://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_Internet\\_connection\\_speeds](https://en.wikipedia.org/wiki/List_of_countries_by_Internet_connection_speeds)



18

## Computer-based texts

<b>Introduction .....</b>	370
<b>Item rendering.....</b>	370
<b>Translation and online item review.....</b>	370
<b>School computer requirements .....</b>	371
<b>System diagnostic.....</b>	371
<b>Test delivery system.....</b>	371
<b>Data capture and scoring student responses .....</b>	373
<b>Open ended coding system.....</b>	373



## INTRODUCTION

In PISA 2015, for the first time, the primary mode of assessment of student skills was computer based. While paper and pencil was an option, the majority of countries chose to implement the entire survey with the computer. All domains were delivered via computer, including the optional financial literacy assessment. There were a total of 90 language versions for the countries using computer based assessment.

This chapter focuses on the functionality and technical implementation of the computer based assessments. It also details the PISA student delivery system (SDS), integrated the assessments with the computer based questionnaires for delivery of the PISA survey in schools. Finally, we conclude with a discussion of the open-ended coding system (OECS), used for coding of open responses.

## ITEM RENDERING

The items for PISA 2015 were implemented using web based technologies: HTML, CSS and JavaScript®. Modern web browsers (such as Firefox v22) provide a rich set of functionalities for attractive presentations and engaging interactivity. At the beginning of the development work, an overall user interface was designed, with common elements such as navigation, help and progress indicators. The items were built in such a way that these common elements were shared, so that the same version was used in all items in each language version.

PISA items are grouped into units consisting of a set of items with a common stimulus. Each unit was constructed independently, with the questions and stimulus components developed first in English, then translated into French to create the two source language versions. Each unit could be viewed on its own, or grouped into test forms for delivery to students as part of the assessments.

In some cases, such as collaborative problem solving (CPS) and the interactive scientific literacy units, common functionalities were split out into shared programming libraries that could be reused in multiple units. For example, in the CPS units, the interactive chat functionality was built as a shared library. For each unit, an XML representation of the chat structure was read in. The library then displayed the chat entries to the student, presented response options, and managed the interactive displays that were unique to the units (shown on the right side of the screen). The library also managed the recording of data and scoring of the student's performance based on unit specific criteria.

As well as the visual aspects of the PISA items, the automated coding of student responses was also implemented using JavaScript®. Shared libraries were created to implement this coding in a common way. The libraries targeted the various response modes used within PISA:

- Form: for all responses using common web form elements such as radio buttons, checkboxes, dropdown menus and textboxes.
- Drag and Drop: for items using drag and drop as the response mode.
- Selection: for items where the response is given by clicking on an object or region of the screen. This can be, for instance, clicking on part of an image, a cell in a table or a segment of text.
- Ad hoc: A general catch all that uses custom JavaScript® code to implement the coding. This was used for unique situations, such as coding for collaborative problem solving and interactive scientific literacy items.

In all cases except the ad hoc coding, the coding for a specific item was specified using rules composed of conditional expressions and Boolean operators. Each library implemented appropriate conditional expressions (e.g., a CONTAINS operator in the Drag and Drop library to test if a drop target held a particular drag element).

## TRANSLATION AND ONLINE ITEM REVIEW

Given the need to support up to 90 versions of each unit, one for each national language, automated support for integration of national translations and adaptations was critical. Supporting this process started when the units were initially developed. The HTML files that implement the display of the unit contain only HTML markup and the text to be shown on the screen. Layout and formatting specifications are stored separately in CSS stylesheets. The text of the units is then extracted from these HTML files and saved to XLIFF (<http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html>) files. XLIFF is a standard file format used for computer supported translation. Once a translation is completed, the XLIFF file is injected into the original source version of the item, resulting in HTML files with the translated text of the unit.



Experience has shown that the quality of a translation is improved if the translators can view their translation in the context of the items they are translating. In an ideal world, translators would work in a completely WYSIWYG (what you see is what you get) mode, so that they are translating directly in the items. This is not technically feasible, and also may tempt translators to focus more on the visual aspects of the items, which are tightly controlled for comparability, rather than the translated text. A good compromise is to provide an easy to use preview capability, giving users a fast way to view their translations as functioning items. For PISA 2015, this capability was provided through an online portal. Users were able to upload an XLIFF file, either partially or completely translated, and in a matter of seconds they received a preview of the given unit in exactly the same view as a student would receive. From this preview, they could interact with the units in the same way as a student responding to the unit, an important factor for the complex units of collaborative problem solving and interactive science. This preview also allowed countries to test and identify potential problems with their translated units before receiving the final software to be used in schools. Problems were fixed as early in the schedule as possible.

## SCHOOL COMPUTER REQUIREMENTS

The goal for PISA 2015 was to use the computers available in the sampled schools with no modifications. The PISA system supported Windows based computers in the field trial and both Windows and Macintosh computers for the main study. The following minimum technical requirements were established for both the field trial and main study.

- CPU speed: At least 1500 MHz
- operating systems: Windows XP® or later, MacOS X® 10.7 or later
- installed memory: 512MB for Windows XP®, 1024MB for newer Windows versions, 2048MB for Macintosh
- free memory: 384MB for Windows XP®, 717MB for newer Windows versions and Macintosh
- screen resolution: 1024 x 768 pixels or greater.

These were the minimum requirements. Computers with higher capabilities would obviously perform better (e.g., respond faster) when delivering the survey, but these were the minimum settings that would provide adequate performance.

## SYSTEM DIAGNOSTIC

In order to verify that the available school computers met these minimum requirements, a system diagnostics application was provided to countries. This application simulated the test delivery system, but rather than showing the PISA tests or questionnaires, it ran a program to check the computer's hardware and software setup and report this back to the user, typically the test administrator or technical support staff in the school. The system diagnostics was provided to countries approximately six months prior to the start of the field trial and main study. This allowed countries to provide the software to sampled schools to determine if their computers could run the PISA survey. Additionally, it was recommended that test administrators run the system diagnostics on the day of the test.

For cases where schools did not have adequate quality or quantity of computers, test administrators brought laptops into the school to augment the available infrastructure. In a few cases, countries chose to administer the PISA tests in all sampled schools on laptops brought into the schools. This avoided "surprises" on the day of the test, where computers were not available or not functioning properly.

## TEST DELIVERY SYSTEM

The PISA 2015 test delivery system, called the student delivery system or SDS, integrated the PISA computer-based assessments and questionnaires for a country, along with a number of components packaged together to run as a standalone application on a USB drive. The SDS did not require network connectivity or external resources to operate. All software and data was on the USB drive, and results were saved back to the USB drive. The SDS could also be run from the computer's hard drive if desired. The components which made up the SDS included:

- Apache web server (<https://httpd.apache.org/>)
- MySQL database engine (<https://www.mysql.com/>)
- PHP interpreter and libraries (<http://php.net/>)
- Firefox Portable web browser ([http://portableapps.com/apps/internet/firefox\\_portable](http://portableapps.com/apps/internet/firefox_portable)).



Together with these open source applications, the actual test and questionnaire content were included. To display this content to the students and collect the results, the PISA test delivery system was implemented. Using portions of the open source TAO test delivery system (<http://www.taotesting.com/>) as a basis, the system was custom built for the needs of PISA. This included implementation of the test flow, which assigns the designated test form and questionnaires to a student, then sequences through the test units and questionnaires in the appropriate order. It also includes the functionality for collecting the survey results and exporting them when the tests are completed. The PISA test delivery system was built primarily using PHP, with JavaScript used for interactive displays and communicating between the web browser and web server.<sup>1</sup>

The system was launched by running a custom executable program (PISAMenu) written for controlling the PISA tests. Separate programs were written for Windows and Macintosh. From this program, a test administrator could launch the PISA tests, launch the system diagnostics, or manage exported data files. These exported files are described below. Launching either the PISA tests or system diagnostics would start the web and database servers, then launch the Firefox web browser to begin the process. When the PISAMenu program was shut down, the various components of the SDS were also terminated.

The Firefox browser used for the PISA tests was configured to run in “kiosk mode”, so that it filled the full screen of the computer, making it difficult for users to access external applications when running the PISA test mode. A keyboard filter was also installed so that students could not easily leave or terminate the browser window, e.g., by pressing Alt-Tab, and switch to another program during the test. The keyboard filter did not completely block such attempts, though. It was not possible to block the Ctrl-Alt-Delete sequence under Windows, as this required installation of a custom software driver at the system level. Our goal was not to install any software on the school computers, so this driver was not used. It was expected that the test administrator would monitor the students during the test and watch for cases of students trying to break out of the system.

The first screen a student would see after the test was started was the option to select one of three sessions: PISA Tests, PISA Questionnaires and Financial Matters. The latter was for financial literacy, and was only shown in countries participating in this international option. After selecting the appropriate session (which usually was done by the test administrator before the students arrived), the student was prompted for a login ID and password. The login ID was the 15 digit student ID assigned by KeyQuest as part of the sampling process. The password was also assigned by KeyQuest and was an 11 digit number. The first five digits comprised a checksum of the student ID, guarding against input errors. The next three digits encoded the test form which should be used for the student. The last three digits were a checksum of the three digit test form number.

After logging in, the student could optionally be shown a screen asking to select a language for this session. While the SDS was built with all the national languages available for a given country, it could be configured to support only one language. This was the recommended method of operation, where the test administrator chose the language configuration when starting the SDS, based on the school where the testing occurred. However, in some countries, it was necessary to allow the students to choose the language of assessment. The typical reason for allowing student choice for the language was for countries and schools with mixed language environments. In these cases, students decided at the time they started the survey session which language to use. The test administrator would guide students through the login and language selection process.

An important facet of the USB setup was protecting the test content on the USB drives. The PISA tests contain secure test materials, and people who obtain a USB drive should not have access to the test items except during the administration of the survey. To accomplish this, the files for rendering all test materials were stored in the MySQL database on each USB drive. The files were stored in an encrypted format, and access to these was controlled via the web server. When a testing session was first started, the PISAMenu program would prompt for the password used to encrypt the files. Each country was assigned a unique password. This password was validated against known encrypted content in the database and then saved for the duration of the testing session. When a request was made to the web server for some part of the test content (e.g., one of the web pages or graphic images), the web server retrieved the content from the database and decrypted it on the fly.

One advantage of the SDS architecture used in 2015 was that it could be run without administrator rights to the local computer. This was a big improvement over past PISA cycles, reducing greatly the amount of technical support needed within the schools.



## DATA CAPTURE AND SCORING STUDENT RESPONSES

Results from the PISA tests and questionnaires were stored on the USB drives. Data was saved as the students answered each question, then exported at key intervals during the survey. At the end of a session, the results from that session were exported in a single password protected ZIP file. For the PISA tests in Session 1 and 3 (the standard PISA domains and the optional financial literacy domain, respectively), the ZIP files contained XML formatted data including logs of the students' actions going through the tests and files with the "variables" exported from the test. The following set of variables were exported for each item in the tests:

- Response: A string representing the raw student response.
- Scored Response: The code assigned to the response when the item was coded automatically.
- Number of Actions: The number of actions taken by the student during the course of interacting with the item. Actions counted were clicks, double-clicks, key presses and drag/drop events.
- Total Time: The total time spent on the item by the student.
- Time to First Action: The time between the first showing of the item and the first action recorded by the system for the item.

Besides these five standard variables, some more complex items, such as the science simulations and collaborative problem solving, had custom variables that were of interest to the test development and psychometric teams. For instance, for the science simulations, the system exported counts of the number of experiments performed and the final set of results from each of these experiments.

An important task in PISA is coding of student responses. For computer delivered tests, many of the item responses can be coded automatically. In PISA 2015, this included multiple choice items, drag and drop items, numeric response items, and complex responses to science simulations. Additionally, in collaborative problem solving, all coding was done automatically, based on the path taken by the student in the chat, as well as other inputs depending on the scenario.

For standard response modes, such as multiple choice or numeric entry, automated coding was done using a rule based system. The correct answer (or partially correct answers in the case of partial credit items) were defined based on Boolean rules defined in a custom syntax. Simple conditionals were possible, e.g., to support different combinations of checkboxes in a multiple selection item where two out of three correct options should be selected. For numeric response items, the rules could check for string matches, which required an exact match against a known correct answer, or numeric matches, which used numeric equivalence to check an answer. For numeric equivalence, for instance, 34.0 would match 34, but they would not match when using string matching.

A challenging part of evaluating numeric responses in an international context like PISA is how to parse the string of characters typed by the student and interpret it as a number. There are differences in decimal and thousands separators that must be taken into account, based on conventions used within countries and local usage. Use of these separators is not always consistent within a country, especially with increased migration and the pervasiveness of the Internet. For PISA 2015, the coding rules tried multiple interpretations of the student response to see if one of them could be coded as correct. The numbers were parsed in different ways, changing the decimal and thousands separators, testing each option to see if a correct response could be granted full or partial credit. Only if no alternate interpretation of the response resulted in a correct answer would the answer be coded as incorrect.

## OPEN ENDED CODING SYSTEM

While automatically coded items formed a significant portion of the tests for PISA 2015, approximately 30% of the items resulted in an open ended response that needed to be coded by a human scorer or coder. On paper, this would be done directly in the test booklets. On the computer, a procedure was necessary to extract the responses provided by the students and present them to human coders. It was important to present these responses in a way that reflected the students' intent. This task is complicated by the fact that these responses could be more than just text. For some items, a student would select an option from a multiple choice part, then type in an explanation for why they chose that option. Additionally, in mathematics, students could use an equation editor to insert complex mathematics notation into their response.

For PISA 2015, the coding of these responses was done using the open ended coding system (OECS). The OECS took the open ended response data generated from the SDS, and, following a coding design specified by the psychometrists,

assigned these responses to coders. Each coder received responses organised by item, so that coders focused on one item at a time. These responses were formatted and saved in PDF files. The PDFs used PDF Form technology ([https://en.wikipedia.org/wiki/Portable\\_Document\\_Format#AcroForms](https://en.wikipedia.org/wiki/Portable_Document_Format#AcroForms)), allowing a coder to record his/her judgement of the student performance. The coded PDF files were then imported back into the OECS and saved in a database. From there, reports were generated with statistics evaluating the reliability of the coding, as well as the completion status for each coder and item. Finally, the coded results were exported and then imported into the Data Management Expert to be integrated with the other PISA data.

The use of PDF files for presentation of response data and collection of the resulting codes had advantages and disadvantages. One advantage was that coders were able to work offline, without need for an Internet connection. But disadvantages included slow PDF generation times when creating the files for the coders, and challenges with managing a large number of PDF files. For some coding designs, there could be hundreds of PDF files for each domain processed during a coding session. Proper managing and accounting of the many files presented a challenge. The OECS organised the files such that a folder of files could be copied for each coder, but coders' computer skills varied and sometimes coded files could be overwritten with files for other items. In the future, if PDF files continue to be used, a user friendly application could be built to help with the management of the files. Alternatively, an online, web-based delivery of the student responses could be developed, obviating the need for files.

### Note

1. The software that implemented the PISA tests in 2015 will be released as open source. Details on availability are not finalised at the time of writing.



---

19

# International data products

<b>Public use files</b> .....	376
<b>Codebooks for the PISA 2015 public use data files</b> .....	377
<b>Data compendia tables</b> .....	378
<b>Data analysis and software tools</b> .....	378
<b>International Database Analyzer</b> .....	380
<b>Population and quality check of the PISA Data Explorer</b> .....	381

Following the data processing and data analysis, data products were delivered to the OECD. These included public use data files and codebooks, compendia tables, and the PISA Data Explorer, a data analysis tool. These data products are available on the OECD website (<http://www.oecd.org/pisa/>). The IEA IDB Analyzer was configured to work with PISA data and can be downloaded at <http://www.iea.nl/our-data>.

## PUBLIC USE FILES

The international public use data files combine all international reportable countries into one file and include an approved set of international variables that are common to all countries. Each national database includes approximately 3 000 common variables for student cognitive and background questionnaire assessments and approximately 600 school and teacher variables. A subset of these were included in the public use data files, made available on the OECD website at <http://www.oecd.org/pisa/>.

## Variables excluded or suppressed for some or all countries

The public use data files include only a subset of the information available in the master databases available to each country. The public use data files do not include any data collected using national adaptations and extensions. Rather, they include only data that were collected or derived across all countries. Further, a sizable number of variables were excluded in consultation with the OECD Secretariat because they i) have little or no analytical utility, ii) were intended for internal or interim purposes only, iii) relate to secure item material, or iv) include personally identifiable data, or at least data that may increase the risk of unintended or indirect disclosure.

The groups of variables excluded from the public use data files are:

1. direct, indirect, and operational identifiers for respondents
2. certain background questionnaire (BQ) or process variables that are available (e.g. country and language), especially detailed free-text entry items
3. all national adaptations and extensions in the BQ
4. original scale score values (theta) before standardisation to an international metric.

As discussed in Chapter 10, countries were given the option of suppressing variables in the public use files. Suppression of variables was approved when data presented a risk to student, school, and/or teacher anonymity, or for technical errors that could not be resolved by data contractors. Suppressed data are represented in the database by means of missing codes.

## File names and content

There are five public use data files: the student questionnaire data file (which also includes estimates of student performance and parent-questionnaire data), the school questionnaire data file, the teacher questionnaire data file, the cognitive item data file and a file with questionnaire timing data. These files include countries/economies/subregions that fully met adjudication criteria. An additional data file contains the data for countries with adjudication issues.

Data files are provided for both SAS and SPSS formats. The files include:

- **Student questionnaire data file (PUF\_COMBINED\_CMB\_STU\_QQQ.zip):** This file includes ID variables, all student questionnaire data (from the Student Background Questionnaire, Educational Career Questionnaire, and Information and Communication Technology Questionnaire), parent-questionnaire data, student and parent-questionnaire scale and derived variables, plausible values (reading, math, and science), and overall and replicate weights.
- **School questionnaire data file (PUF\_COMBINED\_CMB\_SCH\_QQQ.zip):** The school questionnaire data file includes ID variables, school questionnaire data, school questionnaire scale and derived variables, and an overall school weight.
- **Teacher questionnaire data file (PUF\_COMBINED\_CMB\_TCH\_QQQ.zip):** The teacher questionnaire data file includes ID variables, teacher questionnaire data, and teacher questionnaire scale and derived variables.
- **Cognitive Item data file (PUF\_COMBINED\_CMB\_STU\_COG.zip):** The cognitive data file includes ID variables, raw and coded items, computer-based assessment (CBA) item log data (total time and number of actions), as well as some additional CBA cognitive new science information.



- **Questionnaire timing data file (PUF\_COMBINED\_CMB\_STU\_QTM.zip):** The questionnaire timing data file includes CBA questionnaire log data (i.e., total time on a unit/screen).
- **Additional data files for Albania, Argentina, Kazakhstan and Malaysia (PUF\_COMBINED\_CM2\_STU\_QQQ\_COG\_QTM\_SCH\_TCH.zip):** These files include all data for Argentina, Kazakhstan and Malaysia, and student questionnaire data for Albania. Due to issues identified during data adjudication, caution is required when these data. For further information, see Annex A4 of *PISA 2015 Results (Volume I): Excellence and Equity in Education* (OECD, 2016).

Data for student questionnaire items ST016 and ST038 are made available in the *PISA 2015 Results Volume III*, published in April 2017. Financial literacy datasets are available in the *PISA 2015 Results Volume IV*, published in May 2017. Collaborative problem solving datasets are available in the *PISA 2015 Results Volume V*, published in November 2017.

## **Variables used in sampling, weighting and merging**

The variable *STRATUM* is included to differentiate sampling strata. The variable is created as a concatenation of a three-letter country code, a two-digit region identifier and a two-digit original stratum identifier.

The variable *SENGWT* is a normalised (senate) weight variable for analyses of student performance across a group of countries where contributions from each of the countries in the analysis are desired to be equal regardless of their population or sample size. The senate weight makes the population of each country to be 5 000 to ensure an equal contribution by each of the countries in the analysis. This weight is only applicable to the student variables that do not contain missing values. Its application to other variables might be compromised by its dependence on the patterns of missing data.

The student and teacher data files can be merged to the school data file using the variable *CNTSCHID*. *CNTSCHID* is the combination of the three-digit country code and a randomised five-digit number, making it unique across all countries. *CNTSCHID*, *CNTSTUID* (in the student file), and *CNTTCHID* (in the teacher file) have had their values randomised from the original order received during country submission while still retaining the original student to school and teacher to school connection.

## **Missing code conventions**

The data may include up to five MISSING categories:

1. Missing/blank – In the cognitive data, it is used to indicate the respondent was not presented the question according to the survey design or ended the assessment early and did not see the question. In the questionnaire data, it is only used to indicate that the respondent ended the assessment early or despite the opportunity, did not take the questionnaire.
2. No response/omit – The respondent had an opportunity to answer the question but did not respond.
3. Invalid – Used to indicate a questionnaire item was suppressed by country request or that an answer was not conforming to the expected response. For a paper-based questionnaire, the respondent indicated more than one choice for an exclusive-choice question. For a computer-based questionnaire, the response was not in an acceptable range of responses, e.g., the response to a question asking for a percentage was greater than 100.
4. Not applicable – A response was provided even though the response to an earlier question should have directed the respondent to skip that question, or the response could not be determined due to a printing problem or torn booklet. In the questionnaire data, it is also used to indicate missing by design (i.e. the respondent was never given the opportunity to see this question).
5. Valid skip – The question was not answered because a response to an earlier question directed the respondent to skip the question. This code is assigned by Core 3 during data processing.

## **CODEBOOKS FOR THE PISA 2015 PUBLIC USE DATA FILES**

Included with the PISA 2015 main survey data products is a set of data codebooks in Excel format. The data codebook is a printable report containing descriptive information for each variable contained in a corresponding data file. The codebooks report frequencies and percentages for all variables that employ a value scheme for cognitive and questionnaire variables, as well as those that have been derived and/or added during data cleaning. The codebooks are available on the OECD website (<http://www.oecd.org/pisa/>).

The information is displayed with variable names, variable labels, values and value labels. Other metadata are provided, such as variable type (e.g., string or numeric) as well as precision/format. Additionally, the codebooks contain a range of values (minimum and maximum) for those numeric variables that do not employ a value scheme.

Codebooks for the main files are contained in five separate worksheets (**Codebook\_CMB.xlsx**):

1. Student – Student questionnaire data include Parent, Educational Career, and Information Communication and Technology questionnaire data
2. School – School questionnaire data
3. Cognitive – Student cognitive data for reading, mathematics, and science
4. Timing – Student questionnaire timing data
5. Teacher – Teacher questionnaire data.

Codebooks for the additional files for Albania, Argentina, Kazakhstan and Malaysia are contained in a similar set of worksheets in the file **Codebook\_CMS.xlsx**.

## DATA COMPENDIA TABLES

Using the public use files as the source data, the compendia are sets of tables that provide categorical percentages for both cognitive and background items. The compendia support public use file users so that they can gain knowledge of the contents of the data files and use the compendia results so that they are performing public use file analyses correctly. The compendia are available on the OECD website (<http://www.oecd.org/pisa/>).

Questionnaire compendia provide the distribution of students according to the variables collected through the questionnaires. Cognitive compendia provide the distribution of student responses for each test item. Results are provided in Excel format, separately for background questions and test items, and are further broken out by type of questionnaire and by domain (and by gender for cognitive tables). Each Excel file contains multiple worksheets, with each worksheet corresponding to a single variable. The first worksheet in each file is a table of contents that contains a hyperlink to each variable so users can see at a glance which variables are available and can click to go directly to the desired data.

For each questionnaire (EC, ICT, Parent, School, and Student), the percentage of responses in each category are provided in the Excel files with “overall” in the name. Average scale scores corresponding to each category are provided in the files identified by the domains “math”, “read”, and “scie”. The file “**pisa\_bq\_continuous\_overall\_compendium.xls**” provides percentage and percentile data for continuous background variables across all questionnaires. All statistics are calculated using weighted data, with their corresponding standard errors taking into account sampling and measurement uncertainty. The OECD average is created from the 35 current OECD member countries.

The nine Excel files for the cognitive data provide percentages in each response category for the test items. Results are provided separately for females, males, and overall (total) for each domain.

## DATA ANALYSIS AND SOFTWARE TOOLS

Standard analytical packages for the social sciences and educational research do not readily recognise or support handling the complex PISA sample and assessment design. This gap is filled by the two software tools made available to assist database users to access and analyse PISA data and produce basic outputs: the PISA Data Explorer (PDX) and a micro-data analyser. Each of these two software tools addresses a slightly different set of needs. While the PDX is a web-based application that allows relatively easy and publication-ready access to basic estimates of means, totals and proportions, the IEA's IDB Analyzer used in conjunction with the PUFs allows unit record access to the public use database and the opportunity to conduct analysis offline, derive additional variables, and produce various estimates for further use and reporting. The PDX and IEA's IDB Analyzer are described in turn in the remainder of this chapter.

### PISA Data Explorer (PDX)

The PDX is a web-based application that allows the user to query an OECD hosted, secure, PISA International Database via a web browser. In addition to PISA 2015 micro-data, the PDX database contains previous cycle PISA international micro-data that was released in public use files. The PDX is available on the OECD website (<http://www.oecd.org/pisa/>) and provides access to a secure PISA database (protected by the OECD firewalls and security mechanisms) to navigate, analyse, and produce report quality tables and graphics.



The database underlying the PDX is populated using the public use files to import more than 2.4 million unique student records across six PISA cycles. About 5,000 variables across six assessment cycles and over 100 countries and adjudicated subregions are available for analysis. Because certain variables that are included in the public use file (PUF) for secondary analysis are not informative as part of the PDX, they are not included in the PDX database. The majority of variables included only in the PUF relate to the individual cognitive item scores and process information.

The PDX can be used to compute a diverse range of statistics including, but not limited to, means, standard deviations, standard errors, percentages by subgroup, percentages by performance levels and percentiles. All statistics are computed taking into account the sampling and assessment design. In addition, the PDX has the capability of conducting significance testing between statistics from different groups and displaying the results in graphical form. Results from the PDX can be directly exported and saved in Microsoft Word, Microsoft Excel and HTML formats.

Because it is web-based, and processing takes place on a central server, the PDX can be accessed and used with computers that meet fairly simple requirements. The user's computer is used only to create a request or data query, deliver the request to a central server where processing takes place, and then receive and display back the results in a user friendly format.

A typical query consists of the user selecting the domain(s), jurisdiction(s), and variable(s) of interest. Then the user proceeds to select the statistics of interest and format the table. Statistics are calculated for each of the subgroups defined by the variable or variables, for one variable at a time or in cross-tabulation mode. In addition, the user is able to collapse categories for each of these variables and used the collapsed categories in the analysis. All statistics are calculated using weighted data, with their corresponding standard errors taking into account sampling and measurement uncertainty. The user has the option to select whether the standard errors are displayed in the table or not, as well as the precision with which the statistics are displayed. The results can then be displayed in a table or in a graphic.

Regardless of whether the results are displayed in a table or graphic mode, the results can be saved or exported for further post-processing or for inclusion in an external document. Export formats currently available include MS Word, MS Excel, PDF and HTML.

A significance test module allows the user to specify significance testing to be done between subgroup means, percentages and percentiles, within and across cycles, while implementing necessary adjustments that take into account the sample and test design, as well as adjustment for multiple comparisons. Significance test results can be displayed in table or in graphic format.

Table results can be easily exported and manipulated using spreadsheet software, allowing the user to customise the titles and legends of the tables, and to do any required post processing. Likewise, the graphic results can also be exported to be included in documents and used in reports and presentations.

The web application is compatible with many widely used browsers including Internet Explorer 7 and higher, Firefox 3.0 and higher, Google Chrome, and Safari. Target screen resolution is 1024x768. Users should enable JavaScript and pop-ups in their browsers and install Adobe Flash Player 9.0.115 or higher.

### **Import of trend data**

The PISA trend data from 2000 to 2012 were imported into the PDX directly from a database that had been established earlier by the United States Department of Education to develop and support a Data Explorer for PISA and other international studies. These data were taken from all public use files that were available for those cycles and were updated with all subsequent releases of modified or additional data. This approach ensured that all calculated results were consistent with all available OECD reports.

An important outcome of this prior work was the establishment of a naming convention for all data variables to ensure that valid trend comparisons could be made across cycles, even though the variable names as used in the public use file data were not consistent across cycles. This naming convention was extended and applied to all of the 2015 variables in order to ensure continuity and comparability with previous cycles.

In the PISA Data Explorer, the OECD average is created from the 35 current OECD member countries. The same 35 countries are used to create the OECD average for all previous PISA cycles.

### Trend comparison link error factors

Comparisons of performance between two assessments in each domain (e.g., a country's/economy's change in performance between PISA 2000 and PISA 2015 or the change in performance of a subgroup) are calculated using the link error factors shown in Table 19.1.

**Table 19.1 Robust link error for comparisons of performance between PISA 2015 and previous assessments**

Comparison	Mathematics	Reading	Science	Financial literacy
PISA 2000 to 2015		6.8044		
PISA 2003 to 2015	5.6080	5.3907		
PISA 2006 to 2015	3.5111	6.6064	4.4821	
PISA 2009 to 2015	3.7853	3.4301	4.5016	
PISA 2012 to 2015	3.5462	5.2535	3.9228	5.3309

Note: Comparisons between PISA 2015 scores and previous assessments can only be made when the subject first became a major domain. As a result, comparisons in mathematics performance between PISA 2015 and PISA 2000 are not possible, nor are comparisons in science performance between PISA 2015 and PISA 2000, or PISA 2003.

## INTERNATIONAL DATABASE ANALYZER

The IEA International Database Analyzer (IDB Analyzer)<sup>1</sup> is an application developed by the IEA Data Processing and Research Center (IEA-DPC) in Hamburg, Germany, that can be used to analyse data from most major large-scale assessment surveys, including those conducted by OECD, such as PISA. Originally designed for international large-scale assessments, it is also capable of working with national assessments such as the US National Assessment of Educational Progress (NAEP).

The IDB Analyzer creates SPSS or SAS syntax that can be used to perform analysis with these international databases. It generates SPSS or SAS syntax that takes into account information from the sampling design in the computation of sampling variance, and handles the plausible values. The code generated by the IDB Analyzer enables the user to compute descriptive statistics and conduct statistical hypothesis testing among groups in the population without having to write any programming code.

The IDB Analyzer is licensed free of cost, not sold, and is for use only in accordance with the terms of the licensing agreement. While anyone can use the software for free, users do not have ownership of the software itself or its components, including the SPSS and SAS macros, and users are only authorised to use the SPSS and SAS macros in combination with the IDB Analyzer, unless explicitly authorised by the IEA. The software and license expire at the end of each calendar year, when the user will again have to download and reinstall the most current version of the software, and agree to the new license. A complete copy of the licensing agreement is included in the Appendix of the Help Manual of the IDB Analyzer.

The analysis module of the IDB Analyzer provides procedures for the computation of means, percentages, standard deviations, correlations, and regression coefficients for any variable of interest overall for a country, and for specific subgroups within a country. It also computes percentages of people in the population that are within, at, or above benchmarks of performance or within user-defined cut points in the proficiency distribution, percentiles based on the achievement scale, or any other continuous variable.

The analysis module can be used to analyse data files from PISA. The following analyses can be performed with the analysis module:

1. Percentages and means: Computes percentages, means, design effects and standard deviations for selected variables by subgroups defined by the user. The percent of missing responses is included in the output. It also computes t-test statistics of group mean differences taking into account sample dependency.
2. Percentages only: Computes percentages by subgroups defined by the user.
3. Linear regression: Computes linear regression coefficients for selected variables predicting a dependent variable by subgroups defined by the user. The IDB Analyzer has the capability of including plausible values as dependent or independent variables in the linear regression equation. It also has the capability of contrast coding categorical variables (dummy or effect) and including them in the linear regression equation.
4. Logistic regression: Computes logistic regression coefficients for selected variables predicting a dependent dichotomous variable, by subgroups defined by the user. The IDB Analyzer has the capability of including plausible values as independent variables in the logistic regression equation. It also has the capability of contrast coding categorical variables and including them in the logistic regression equation. When used with SAS, the user can also specify multinomial logistic regression models.



5. Benchmarks: Computes percent of the population meeting a set of user-specified performance or achievement benchmarks by subgroups defined by the user. It computes these percentages in two modes: cumulative (percent of the population at or above given points in the distribution) or discrete (percent of the population within given points of the distribution). It can also compute the mean of an analysis variable for those at a particular achievement level when the discrete option is selected. New in 2016 is the computation of group mean and percent differences between groups taking into account sample dependency.
6. Correlations: Computes correlation for selected variables by subgroups defined by the grouping variable(s). The IDB Analyzer is capable of computing the correlation between sets of plausible values.
7. Percentiles: Computes the score points that separate a given proportion of the distribution of scores by subgroups defined by the grouping variable(s).
8. Differences by Performance Groups: Computes the means on an analysis variable by subgroups defined by background variables and performance level. When there are two subgroups within a performance level, it computes significance testing of the difference between these two groups. Currently this functionality is only available with SPSS.

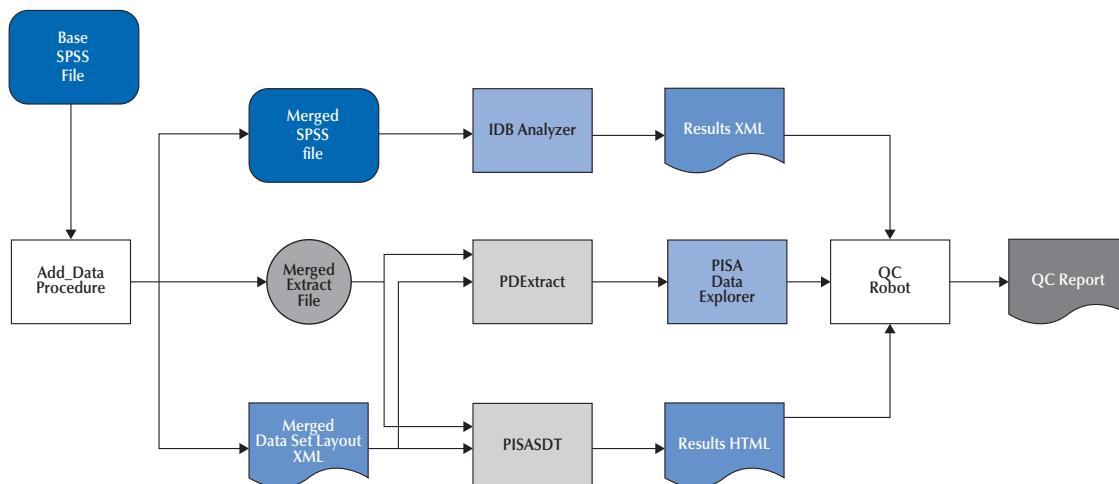
When calculating these statistics, the IDB Analyzer has the capability of using any continuous or categorical variable in the database, or make use of scores in the form of plausible values. When using plausible values, the IDB Analyzer generates SPSS or SAS code that takes into account the multiple imputation methodology in the calculation of the variance for statistics, as it applies to the corresponding study.

All procedures offered within the analysis module of the IDB Analyzer make use of appropriate sampling weights and standard errors of the statistics that are computed according to the variance estimation procedure required by the design as it applies to the corresponding study.

## POPULATION AND QUALITY CHECK OF THE PISA DATA EXPLORER

The process to populate the PISA Data Explorer database and confirm the results it produces is summarised in Figure 19.1 below. This process was applied separately to the data from each country.

■ Figure 19.1 ■  
**PISA database population and quality control**



The Base SPSS file contained the data as forwarded to the appropriate country for its analysis and reporting.

The Add\_Data procedure performed two functions. The first was conditional on whether a country provided supplemental data that was collected or derived and merged these data with the Base file. The second function created two files from the enhanced Base file: an ASCII text rectangular file containing the data values extracted from the Base file and an XML file containing information about the extracted data variables (location, format, labels). This Data Set Layout (DSL) XML is structured in a proprietary ETS schema.

The PDExtract program used the information from an input parameter file to process the data from the Extract file and metadata from the DSL file to produce a series of text files suitable for loading into the appropriate tables in the PISA Data Explorer (PDX) database. The program also produced a SQL script that is customised for performing the loading of these tables and contains a procedure for forming the data tables used by the PDX.

The PISASDT program also used the information from an input parameter file as well as a list of data variable names to calculate and produce summary data tables (SDT) – one analysis for each scale score. Each table in the analysis was a one-way tabulation of various statistics for each category of a given variable. The statistics pertained to a scale score and include percentage, average score and percentages within the benchmark levels. Each statistic was accompanied by the standard error estimate, degrees of freedom, number of cases on which the statistic is based and number of strata on which the standard error was based. All of these results were stored in an HTML document in full precision. This document may be viewed with any of the popular Internet browsers when accompanied by the appropriate Cascading Style Sheet (CSS) document, which ETS provided. The document may also be parsed or translated to produce Excel workbooks and report quality tables, among others.

In the QC Robot procedure, the Results HTML document from the PISASDT program was used to generate analysis requests for the PDX, one for each variable, and the results returned from the PDX were compared with those in the HTML document. The results of these comparisons were posted to the QC Report document where differences above specified criteria were flagged and subsequently examined.

The only statistics that can be reported in the PDX which cannot be calculated by the PISASDT program are the percentiles. Because the calculation of the percentiles within the PDX uses more resources than the other statistics, only a subset of critical variables was selected for quality-assurance analysis. The Analyzer reads data from the Base SPSS file, uses SPSS macros to calculate the desired percentile statistics, and writes the results to an XML file. The QC Robot procedure processed this XML file in the same way as the HTML file from the PISASDT program and added the comparison results to the QC Report file.

Prior to the first execution of the procedure described above, the Analyzer and the PISASDT programs were extensively calibrated with each other to ensure that the Merged SPSS and Merged Extract files were isomorphic and produced identical results for the statistics common to both programs.

### Note

1. <http://www.iea.nl/our-data>.

### Reference

OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, OECD Publishing, Paris, <http://dx.doi.org/10.1787/9789264266490-en>.



# Annexes

<b>Annex A</b>	Main survey item pool classification .....	384
<b>Annex B</b>	Contrast coding used in conditioning.....	427
<b>Annex C</b>	Standard errors of means, sample sizes, school variance estimates, and other sampling outcomes .....	428
<b>Annex D</b>	Mapping of ISCED to years.....	435
<b>Annex E</b>	National household possession items .....	436
<b>Annex F</b>	Technical standards for PISA 2015 .....	438
<b>Annex G</b>	Common and unique item parameters in each domain, by countries and languages .....	454
<b>Annex H</b>	Scalar or metric invariant trend items in each domain .....	455
<b>Annex I</b>	PISA contractors, staff and consultants .....	456

Note regarding B-S-J-G (China)

B-S-J-G (China) refers to the four PISA participating China provinces : Beijing, Shanghai, Jiangsu, Guangdong.

Note regarding CABA (Argentina)

CABA (Argentina) refers to the Ciudad Autónoma de Buenos Aires, Argentina.

Note regarding FYROM

FYROM refers to the Former Yugoslav Republic of Macedonia.

Notes regarding Cyprus

Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Note regarding Israel

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.



## ANNEX A – MAIN SURVEY ITEM POOL CLASSIFICATION

[Part 1/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	CBA item ID in main survey analysis output	PBA item ID in main survey analysis output	Unit name	Mode [paper-based (PB); computer-based (CB)]	2015 field trial and main survey cluster	Sequence in cluster	Sequence in unit
S131Q02	DS131Q02C	PS131Q02	Good Vibrations	PB and CB	S03	4	1/2
S131Q04	DS131Q04C	PS131Q04	Good Vibrations	PB and CB	S03	5	2/2
S252Q01	CS252Q01S	PS252Q01S	South Rainea	PB and CB	S06	6	1/3
S252Q02	CS252Q02S	PS252Q02S	South Rainea	PB and CB	S06	7	2/3
S252Q03	CS252Q03S	PS252Q03S	South Rainea	PB and CB	S06	8	3/3
S256Q01	CS256Q01S	PS256Q01S	Spoons	PB and CB	S02, U1	5	1/1
S268Q01	CS268Q01S	PS268Q01S	Algae	PB and CB	S04	8	1/3
S268Q02	DS268Q02C	S268Q02	Algae	PB and CB	S04	9	2/3
S268Q06	CS268Q06S	PS268Q06S	Algae	PB and CB	S04	10	3/3
S269Q01	DS269Q01C	PS269Q01	Earth's Temperature	PB and CB	S01	1	1/3
S269Q03	DS269Q03C	PS269Q03	Earth's Temperature	PB and CB	S01	2	2/3
S269Q04	CS269Q04S	PS269Q04S	Earth's Temperature	PB and CB	S01	3	3/3
S304Q01	DS304Q01C	S304Q01	Water	PB and CB	S05	9	1/4
S304Q02	CS304Q02S	PS304Q02S	Water	PB and CB	S05	10	2/4
S304Q03	DS304Q03aC	S304Q03a	Water	PB and CB	S05	11	3/4
S304Q03	DS304Q03bC	S304Q03b	Water	PB and CB	S05	12	4/4
S326Q01	DS326Q01C	PS326Q01	Milk	PB and CB	S02	1	1/4
S326Q02	DS326Q02C	PS326Q02	Milk	PB and CB	S02	2	2/4
S326Q03	CS326Q03S	PS326Q03S	Milk	PB and CB	S02	3	3/4
S326Q04	CS326Q04S	PS326Q04S	Milk	PB and CB	S02	4	4/4
S327Q01	CS327Q01S	PS327Q01S	Tidal Energy	PB and CB	S06	9	1/2
S408Q01	CS408Q01S	PS408Q01S	Wild Oat Grass	PB and CB	S01	4	1/4
S408Q03	DS408Q03C	PS408Q03	Wild Oat Grass	PB and CB	S01	5	2/4
S408Q04	CS408Q04S	PS408Q04S	Wild Oat Grass	PB and CB	S01	6	3/4
S408Q05	CS408Q05S	PS408Q05S	Wild Oat Grass	PB and CB	S01	7	4/4
S413Q04	CS413Q04S	PS413Q04S	Plastic Age	PB and CB	S02	10	2/3
S413Q05	CS413Q05S	PS413Q05S	Plastic Age	PB and CB	S02	11	3/3
S413Q06	CS413Q06S	PS413Q06	Plastic Age	PB and CB	S02	9	1/3
S415Q02	CS415Q02S	PS415Q02S	Solar Power Generation (Solar Panels)	PB and CB	S03, U2	16	2/3
S415Q07	CS415Q07S	PS415Q07S	Solar Power Generation (Solar Panels)	PB and CB	S03, U2	15	1/3
S415Q08	CS415Q08S	PS415Q08S	Solar Power Generation (Solar Panels)	PB and CB	S03, U2	17	3/3
S416Q01	DS416Q01C	S416Q01	The Moon	PB and CB	S05	13	1/1
S421Q01	CS421Q01S	S421Q01	Big and Small	PB and CB	S06, U2	3	1/3
S421Q02	CS421Q02S	S421Q02	Big and Small	PB and CB	S06, U2	4	2/3
S421Q03	CS421Q03S	S421Q03	Big and Small	PB and CB	S06, U2	5	3/3
S425Q02	CS425Q02S	PS425Q02S	Penguin Island	PB and CB	S02	17	3/4
S425Q03	DS425Q03C	PS425Q03	Penguin Island	PB and CB	S02	15	1/4
S425Q04	DS425Q04C	PS425Q04	Penguin Island	PB and CB	S02	18	4/4
S425Q05	CS425Q05S	PS425Q05S	Penguin Island	PB and CB	S02	16	2/4
S428Q01	CS428Q01S	PS428Q01S	Bacteria in Milk	PB and CB	S03, U1	6	1/3
S428Q03	CS428Q03S	PS428Q03S	Bacteria in Milk	PB and CB	S03, U1	7	2/3
S428Q05	DS428Q05C	PS428Q05	Bacteria in Milk	PB and CB	S03, U1	8	3/3



[Part 2/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	CBA item ID in main survey analysis output	PBA item ID in main survey analysis output	Unit name	Mode [paper-based (PB); computer-based (CB)]	2015 field trial and main survey cluster	Sequence in cluster	Sequence in unit
S437Q01	CS437Q01S	PS437Q01S	Extinguishing Fires	PB and CB	S05	5	1/4
S437Q03	CS437Q03S	PS437Q03S	Extinguishing Fires	PB and CB	S05	6	2/4
S437Q04	CS437Q04S	PS437Q04S	Extinguishing Fires	PB and CB	S05	7	3/4
S437Q06	DS437Q06C	S437Q06	Extinguishing Fires	PB and CB	S05	8	4/4
S438Q01	CS438Q01S	PS438Q01S	Green Parks	PB and CB	S03	12	1/3
S438Q02	CS438Q02S	PS438Q02S	Green Parks	PB and CB	S03	13	2/3
S438Q03	DS438Q03C	PS438Q03	Green Parks	PB and CB	S03	14	3/3
S458Q01	DS458Q01C	S458Q01	The Ice Mummy	PB and CB	S06	1	1/2
S458Q02	CS458Q02S	PS458Q02S	The Ice Mummy	PB and CB	S06	2	2/2
S465Q01	DS465Q01C	PS465Q01	Different Climates	PB and CB	S03	1	1/3
S465Q02	CS465Q02S	PS465Q02S	Different Climates	PB and CB	S03	2	2/3
S465Q04	CS465Q04S	PS465Q04S	Different Climates	PB and CB	S03	3	3/3
S466Q01	CS466Q01S	PS466Q01S	Forest Fires	PB and CB	S01, U1	16	1/3
S466Q05	CS466Q05S	PS466Q05S	Forest Fires	PB and CB	S01, U1	18	3/3
S466Q07	CS466Q07S	PS466Q07S	Forest Fires	PB and CB	S01, U1	17	2/3
S476Q01	CS476Q01S	PS476Q01S	Heart Surgery	PB and CB	S04, U2	1	1/3
S476Q02	CS476Q02S	PS476Q02S	Heart Surgery	PB and CB	S04, U2	2	2/3
S476Q03	CS476Q03S	PS476Q03S	Heart Surgery	PB and CB	S04, U2	3	3/3
S478Q01	CS478Q01S	PS478Q01S	Antibiotics	PB and CB	S02	6	1/3
S478Q02	CS478Q02S	PS478Q02S	Antibiotics	PB and CB	S02	7	2/3
S478Q03	CS478Q03S	PS478Q03S	Antibiotics	PB and CB	S02	8	3/3
S495Q01	CS495Q01S	PS495Q01S	Radiotherapy	PB and CB	S04	5	2/4
S495Q02	CS495Q02S	PS495Q02S	Radiotherapy	PB and CB	S04	6	3/4
S495Q03	DS495Q03C	S495Q03	Radiotherapy	PB and CB	S04	7	4/4
S495Q04	CS495Q04S	PS495Q04S	Radiotherapy	PB and CB	S04	4	1/4
S498Q02	CS498Q02S	PS498Q02S	Experimental Digestion	PB and CB	S02	12	1/3
S498Q03	CS498Q03S	PS498Q03S	Experimental Digestion	PB and CB	S02	13	2/3
S498Q04	DS498Q04C	PS498Q04	Experimental Digestion	PB and CB	S02	14	3/3
S510Q01	CS510Q01S	PS510Q01S	Magnetic Hovertrain	PB and CB	S05	3	1/2
S510Q04	DS510Q04C	S510Q04	Magnetic Hovertrain	PB and CB	S05	4	2/2
S514Q02	DS514Q02C	PS514Q02	Development and Disaster	PB and CB	S03	9	1/3
S514Q03	DS514Q03C	PS514Q03	Development and Disaster	PB and CB	S03	10	2/3
S514Q04	DS514Q04C	PS514Q04	Development and Disaster	PB and CB	S03	11	3/3
S519Q01	DS519Q01C	PS519Q01	Airbags	PB and CB	S01	10	1/3
S519Q02	CS519Q02S	PS519Q02S	Airbags	PB and CB	S01	11	2/3
S519Q03	DS519Q03C	PS519Q03	Airbags	PB and CB	S01	12	3/3
S521Q02	CS521Q02S	PS521Q02S	Cooking Outdoors	PB and CB	S01	8	1/2
S521Q06	CS521Q06S	PS521Q06S	Cooking Outdoors	PB and CB	S01	9	2/2
S524Q06	CS524Q06S	PS524Q06S	Penicillin Manufacture	PB and CB	S05	1	1/2
S524Q07	DS524Q07C	S524Q07	Penicillin Manufacture	PB and CB	S05	2	2/2
S527Q01	CS527Q01S	PS527Q01S	Extinction of the Dinosaurs	PB and CB	S01	13	1/3
S527Q03	CS527Q03S	PS527Q03S	Extinction of the Dinosaurs	PB and CB	S01	14	2/3
S527Q04	CS527Q04S	PS527Q04S	Extinction of the Dinosaurs	PB and CB	S01	15	3/3

[Part 3/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	Item format - CBA	Item format - PBA	Context 1 (2015)	Context 1 (2006)	Context 2	Competency (2015)	Competency (2006)
S131Q02	Open Response - Human Coded	Open Response - Human Coded	Personal	Social	Health & Disease	Interpret data and evidence scientifically	Using scientific evidence
S131Q04	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Health & Disease	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S252Q01	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	NA	Natural Resources	Interpret data and evidence scientifically	Missing
S252Q02	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	NA	Natural Resources	Interpret data and evidence scientifically	Missing
S252Q03	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	NA	Natural Resources	Interpret data and evidence scientifically	Missing
S256Q01	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Personal	Personal	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S268Q01	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Hazards	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S268Q02	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Hazards	Explain phenomena scientifically	Explaining phenomena scientifically
S268Q06	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Global	Social	Environmental Quality	Explain phenomena scientifically	Explaining phenomena scientifically
S269Q01	Open Response - Human Coded	Open Response - Human Coded	Global	Global	Environmental Quality	Explain phenomena scientifically	Explaining phenomena scientifically
S269Q03	Open Response - Human Coded	Open Response - Human Coded	Global	Global	Environmental Quality	Explain phenomena scientifically	Explaining phenomena scientifically
S269Q04	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Global	Global	Environmental Quality	Explain phenomena scientifically	Explaining phenomena scientifically
S304Q01	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Natural Resources	Interpret data and evidence scientifically	Using scientific evidence
S304Q02	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Natural Resources	Explain phenomena scientifically	Explaining phenomena scientifically
S304Q03	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Global	Frontiers	Explain phenomena scientifically	Using scientific evidence
S304Q03	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S326Q01	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Personal	Health & Disease	Interpret data and evidence scientifically	Using scientific evidence
S326Q02	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Personal	Health & Disease	Interpret data and evidence scientifically	Using scientific evidence
S326Q03	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Personal	Health & Disease	Interpret data and evidence scientifically	Using scientific evidence
S326Q04	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Personal	Health & Disease	Explain phenomena scientifically	Explaining phenomena scientifically
S327Q01	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Global	NA	Natural Resources	Explain phenomena scientifically	NA
S408Q01	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Natural Resources	Explain phenomena scientifically	Explaining phenomena scientifically
S408Q03	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Natural Resources	Explain phenomena scientifically	Explaining phenomena scientifically
S408Q04	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Natural Resources	Explain phenomena scientifically	Explaining phenomena scientifically
S408Q05	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Natural Resources	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S413Q04	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S413Q05	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S413Q06	Complex Multiple Choice - Computer Scored	Open Response - Human Coded	Local/ National	Personal	Frontiers	Interpret data and evidence scientifically	Explaining phenomena scientifically
S415Q02	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Global	Natural Resources	Explain phenomena scientifically	Explaining phenomena scientifically
S415Q07	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Personal	Natural Resources	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S415Q08	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Global	Global	Natural Resources	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S416Q01	Open Response - Human Coded	Open Response - Human Coded	Global	Global	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S421Q01	Open Response - Computer Scored	Open Response - Human Coded	Personal	Personal	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S421Q02	Open Response - Computer Scored	Open Response - Human Coded	Personal	Personal	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S421Q03	Open Response - Computer Scored	Open Response - Human Coded	Personal	Personal	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S425Q02	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Environmental Quality	Interpret data and evidence scientifically	Using scientific evidence
S425Q03	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Environmental Quality	Explain phenomena scientifically	Explaining phenomena scientifically
S425Q04	Open Response - Human Coded	Open Response - Human Coded	Global	Social	Environmental Quality	Evaluate and design scientific enquiry	Using scientific evidence
S425Q05	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Environmental Quality	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S428Q01	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Health & Disease	Interpret data and evidence scientifically	Using scientific evidence
S428Q03	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Health & Disease	Interpret data and evidence scientifically	Using scientific evidence
S428Q05	Open Response - Human Coded	Open Response - Human Coded	Global	Social	Health & Disease	Explain phenomena scientifically	Explaining phenomena scientifically



[Part 4/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	Item format - CBA	Item format - PBA	Context 1 (2015)	Context 1 (2006)	Context 2	Competency (2015)	Competency (2006)
S437Q01	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Hazards	Explain phenomena scientifically	Explaining phenomena scientifically
S437Q03	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Hazards	Explain phenomena scientifically	Explaining phenomena scientifically
S437Q04	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Hazards	Explain phenomena scientifically	Explaining phenomena scientifically
S437Q06	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Hazards	Explain phenomena scientifically	Explaining phenomena scientifically
S438Q01	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Natural Resources	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S438Q02	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Natural Resources	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S438Q03	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Natural Resources	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S458Q01	Open Response - Human Coded	Open Response - Human Coded	Global	Global	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S458Q02	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Global	Global	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S465Q01	Open Response - Human Coded	Open Response - Human Coded	Global	Global	Natural Resources	Interpret data and evidence scientifically	Using scientific evidence
S465Q02	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Global	Global	Natural Resources	Explain phenomena scientifically	Explaining phenomena scientifically
S465Q04	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Global	Global	Natural Resources	Explain phenomena scientifically	Explaining phenomena scientifically
S466Q01	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Hazards	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S466Q05	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Hazards	Interpret data and evidence scientifically	Using scientific evidence
S466Q07	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Hazards	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S476Q01	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Personal	Health & Disease	Explain phenomena scientifically	Explaining phenomena scientifically
S476Q02	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Personal	Health & Disease	Explain phenomena scientifically	Explaining phenomena scientifically
S476Q03	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Personal	Health & Disease	Explain phenomena scientifically	Explaining phenomena scientifically
S478Q01	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Personal	Health & Disease	Explain phenomena scientifically	Explaining phenomena scientifically
S478Q02	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Personal	Health & Disease	Interpret data and evidence scientifically	Using scientific evidence
S478Q03	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Personal	Health & Disease	Explain phenomena scientifically	Explaining phenomena scientifically
S495Q01	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S495Q02	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S495Q03	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S495Q04	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Frontiers	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S498Q02	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Frontiers	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S498Q03	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Local/ National	Social	Frontiers	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S498Q04	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S510Q01	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S510Q04	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Personal	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S514Q02	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Environmental Quality	Explain phenomena scientifically	Using scientific evidence
S514Q03	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Environmental Quality	Explain phenomena scientifically	Explaining phenomena scientifically
S514Q04	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Environmental Quality	Interpret data and evidence scientifically	Using scientific evidence
S519Q01	Open Response - Human Coded	Open Response - Human Coded	Personal	Social	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S519Q02	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Personal	Social	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S519Q03	Open Response - Human Coded	Open Response - Human Coded	Personal	Social	Frontiers	Evaluate and design scientific enquiry	Identifying scientific questions/issues
S521Q02	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Personal	Personal	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S521Q06	Simple Multiple Choice - Computer Scored	Simple Multiple Choice - Data Entered	Personal	Personal	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S524Q06	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Local/ National	Social	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S524Q07	Open Response - Human Coded	Open Response - Human Coded	Local/ National	Social	Frontiers	Explain phenomena scientifically	Using scientific evidence
S527Q01	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Global	Global	Frontiers	Interpret data and evidence scientifically	Using scientific evidence
S527Q03	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Global	Global	Frontiers	Explain phenomena scientifically	Explaining phenomena scientifically
S527Q04	Complex Multiple Choice - Computer Scored	Complex Multiple Choice - Data Entered	Global	Global	Frontiers	Interpret data and evidence scientifically	Explaining phenomena scientifically

[Part 5/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	Knowledge (2015)	Knowledge (2006)	System (2015)	System (2006)	Depth of knowledge	Unit origin	Language of submission	Source
S131Q02	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Low	ACER	English	2012
S131Q04	Procedural	Knowledge about Science - Scientific Enquiry	Living	NA	Medium	ACER	English	2012
S252Q01	Content	Missing	Earth and Space	NA	Medium	Korea	Korean	2003
S252Q02	Content	Missing	Earth and Space	NA	Medium	Korea	Korean	2003
S252Q03	Procedural	Missing	Earth and Space	NA	Medium	Korea	Korean	2003
S256Q01	Content	Knowledge of Science	Physical	Physical Systems	Low	TIMSS	English	2012
S268Q01	Procedural	Knowledge about Science - Scientific Enquiry	Living	NA	Medium	Australia	English	2006
S268Q02	Content	Knowledge of Science	Living	Living Systems	Medium	Australia	English	2006
S268Q06	Content	Knowledge of Science	Living	Living Systems	Low	Australia	English	2006
S269Q01	Content	Knowledge of Science	Earth and Space	Earth and Space Systems	Low	CITO	Dutch	2012
S269Q03	Content	Knowledge of Science	Living	Living Systems	Medium	CITO	Dutch	2012
S269Q04	Content	Knowledge of Science	Physical	Physical Systems	Low	CITO	Dutch	2012
S304Q01	Content	Knowledge of Science	Physical	Physical Systems	Medium	CITO	Dutch	2006
S304Q02	Content	Knowledge of Science	Physical	Physical Systems	Medium	CITO	Dutch	2006
S304Q03	Content	Knowledge of Science	Physical	Technology Systems	Medium	CITO	Dutch	2006
S304Q03	Content	Knowledge of Science	Physical	Technology Systems	Medium	CITO	Dutch	2006
S326Q01	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Medium	CITO	Dutch	2012
S326Q02	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Medium	CITO	Dutch	2012
S326Q03	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Medium	CITO	Dutch	2012
S326Q04	Content	Knowledge of Science	Living	Living Systems	Low	CITO	Dutch	2012
S327Q01	Content	NA	Earth and Space	NA	Medium	ACER	English	2003
S408Q01	Content	Knowledge of Science	Living	Living Systems	Medium	ILS	Norwegian	2012
S408Q03	Content	Knowledge of Science	Living	Living Systems	High	ILS	Norwegian	2012
S408Q04	Content	Knowledge of Science	Living	Living Systems	Medium	ILS	Norwegian	2012
S408Q05	Procedural	Knowledge about Science - Scientific Enquiry	Living	NA	High	ILS	Norwegian	2012
S413Q04	Content	Knowledge of Science	Physical	Technology Systems	Medium	IPN	German	2012
S413Q05	Content	Knowledge of Science	Physical	Technology Systems	High	IPN	German	2012
S413Q06	Content	Knowledge of Science	Physical	Physical Systems	Medium	IPN	German	2012
S415Q02	Content	Knowledge of Science	Earth and Space	Earth and Space Systems	Low	NIER	Japanese	2012
S415Q07	Epistemic	Knowledge about Science - Scientific Enquiry	Earth and Space	NA	Medium	ACER	English	2012
S415Q08	Epistemic	Knowledge about Science - Scientific Enquiry	Earth and Space	NA	Low	ACER	English	2012
S416Q01	Content	Knowledge about Science - Scientific Explanations	Earth and Space	NA	High	ILS	Norwegian	2006
S421Q01	Content	Knowledge of Science	Physical	Physical Systems	Low	ILS	Norwegian	2006
S421Q02	Content	Knowledge of Science	Living	Living Systems	Low	ILS	Norwegian	2006
S421Q03	Content	Knowledge of Science	Earth and Space	Earth and Space Systems	Low	ILS	Norwegian	2006
S425Q02	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	High	ACER	English	2012
S425Q03	Content	Knowledge of Science	Living	Living Systems	Low	ACER	English	2012
S425Q04	Epistemic	Knowledge about Science - Scientific Enquiry	Living	NA	Medium	ACER	English	2012
S425Q05	Procedural	Knowledge about Science - Scientific Enquiry	Living	Living Systems	Medium	ACER	English	2012
S428Q01	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Low	IPN	German	2012
S428Q03	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Medium	IPN	German	2012
S428Q05	Content	Knowledge of Science	Living	Living Systems	Medium	IPN	German	2012



[Part 6/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	Knowledge (2015)	Knowledge (2006)	System (2015)	System (2006)	Depth of knowledge	Unit origin	Language of submission	Source
S437Q01	Content	Knowledge of Science	Physical	Physical Systems	Low	ACER	English	2006
S437Q03	Content	Knowledge of Science	Physical	Physical Systems	Low	ACER	English	2006
S437Q04	Content	Knowledge of Science	Physical	Physical Systems	Low	ACER	English	2006
S437Q06	Content	Knowledge of Science	Physical	Physical Systems	Medium	ACER	English	2006
S438Q01	Procedural	Knowledge about Science - Scientific Enquiry	Living	NA	Low	ACER	English	2012
S438Q02	Procedural	Knowledge about Science - Scientific Enquiry	Physical	NA	Medium	ACER	English	2012
S438Q03	Epistemic	Knowledge about Science - Scientific Enquiry	Physical	NA	Medium	ACER	English	2012
S458Q01	Content	Knowledge of Science	Living	Living Systems	Medium	ILS	Norwegian	2006
S458Q02	Content	Knowledge of Science	Living	Living Systems	Medium	ILS	Norwegian	2006
S465Q01	Procedural	Knowledge about Science - Scientific Explanations	Earth and Space	NA	Medium	ILS	Norwegian	2012
S465Q02	Content	Knowledge of Science	Earth and Space	Earth and Space Systems	Low	ILS	Norwegian	2012
S465Q04	Content	Knowledge of Science	Earth and Space	Earth and Space Systems	Low	ILS	Norwegian	2012
S466Q01	Procedural	Knowledge about Science - Scientific Enquiry	Physical	NA	Medium	ILS	Norwegian	2012
S466Q05	Procedural	Using scientific evidence	Physical	NA	Medium	ILS	Norwegian	2012
S466Q07	Epistemic	Knowledge about Science - Scientific Enquiry	Physical	NA	Medium	ILS	Norwegian	2012
S476Q01	Content	Knowledge of Science	Living	Living Systems	Low	New Zealand	English	2006
S476Q02	Content	Knowledge of Science	Living	Living Systems	Low	New Zealand	English	2006
S476Q03	Content	Knowledge of Science	Living	Living Systems	Medium	New Zealand	English	2006
S478Q01	Content	Knowledge of Science	Living	Living Systems	Low	France	French	2012
S478Q02	Epistemic	Knowledge about Science - Scientific Explanations	Living	NA	Medium	France	French	2012
S478Q03	Content	Knowledge of Science	Living	Living Systems	Low	France	French	2012
S495Q01	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Medium	France	French	2006
S495Q02	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Medium	France	French	2006
S495Q03	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Medium	France	French	2006
S495Q04	Epistemic	Knowledge about Science - Scientific Enquiry	Living	NA	Low	France	French	2006
S498Q02	Procedural	Knowledge about Science - Scientific Enquiry	Physical	NA	Medium	France	French	2012
S498Q03	Procedural	Knowledge about Science - Scientific Enquiry	Physical	NA	High	France	French	2012
S498Q04	Procedural	Knowledge about Science - Scientific Explanations	Living	NA	Medium	France	French	2012
S510Q01	Content	Knowledge of Science	Physical	Physical Systems	Medium	Belgium	Dutch	2006
S510Q04	Content	Knowledge of Science	Physical	Physical Systems	Medium	Belgium	Dutch	2006
S514Q02	Content	Knowledge of Science	Physical	Technology Systems	Low	NIER	Japanese	2012
S514Q03	Content	Knowledge of Science	Earth and Space	Earth and Space Systems	Medium	NIER	Japanese	2012
S514Q04	Epistemic	Knowledge of Science	Earth and Space	Technology Systems	Medium	NIER	Japanese	2012
S519Q01	Procedural	Knowledge about Science - Scientific Explanations	Physical	NA	High	France	French	2012
S519Q02	Content	Knowledge of Science	Physical	Physical Systems	Low	France	French	2012
S519Q03	Epistemic	Knowledge about Science - Scientific Enquiry	Living	NA	Medium	France	French	2012
S521Q02	Content	Knowledge of Science	Physical	Physical Systems	Low	ACER	English	2012
S521Q06	Content	Knowledge of Science	Physical	Physical Systems	Low	ACER	English	2012
S524Q06	Content	Knowledge of Science	Living	Technology Systems	Low	IPN	German	2006
S524Q07	Content	Knowledge about Science - Scientific Explanations	Living	NA	Medium	IPN	German	2006
S527Q01	Epistemic	Knowledge about Science - Scientific Explanations	Earth and Space	NA	Medium	Korea	Korean	2012
S527Q03	Content	Knowledge of Science	Earth and Space	Earth and Space Systems	Low	Korea	Korean	2012
S527Q04	Content	Knowledge of Science	Earth and Space	Earth and Space Systems	Medium	Korea	Korean	2012

[Part 7/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	CBA international % correct	CBA international % correct S.E.	CBA item parameters (RP = 0.50)				CBA thresholds (RP = .62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
S131Q02	45.02	0.27	1.425	0.075			537		Level 3
S131Q04	26.78	0.24	1.211	0.551			624		Level 4
S252Q01	49.38	0.28	0.858	0.032			553		Level 3
S252Q02	63.78	0.27	0.458	-0.824			458		Level 2
S252Q03	52.79	0.28	0.708	-0.196			526		Level 3
S256Q01	87.63	0.18	1.000	-1.412			302		Level 1b
S268Q01	66.37	0.26	1.000	-0.580			442		Level 2
S268Q02	29.94	0.25	1.571	0.335			578		Level 4
S268Q06	45.11	0.27	0.916	0.088			559		Level 4
S269Q01	44.54	0.26	1.541	0.054			531		Level 3
S269Q03	37.35	0.25	1.625	0.230			559		Level 4
S269Q04	29.24	0.25	0.796	0.637			659		Level 5
S304Q01	31.91	0.25	1.387	0.369			588		Level 4
S304Q02	54.64	0.27	1.000	-0.323			485		Level 3
S304Q03	33.74	0.26	1.588	0.258			565		Level 4
S304Q03	44.28	0.27	2.475	0.009			512		Level 3
S326Q01	48.17	0.27	1.208	-0.194			498		Level 3
S326Q02	54.09	0.28	1.824	-0.206			483		Level 2
S326Q03	52.33	0.27	1.421	-0.144			501		Level 3
S326Q04	23.33	0.23	1.000	0.851			682		Level 5
S327Q01	49.47	0.27	0.991	-0.066			529		Level 3
S408Q01	56.24	0.26	0.965	-0.338			484		Level 2
S408Q03	24.14	0.23	0.983	0.618			644		Level 5
S408Q04	47.99	0.26	0.653	-0.098			548		Level 3
S408Q05	37.75	0.26	0.941	0.227			580		Level 4
S413Q04	36.78	0.26	0.963	0.249			583		Level 4
S413Q05	64.88	0.26	0.908	-0.499			460		Level 2
S413Q06	32.43	0.26	1.623	0.239			561		Level 4
S415Q02	71.67	0.26	1.409	-0.672			412		Level 2
S415Q07	69.27	0.26	0.902	-0.677			431		Level 2
S415Q08	54.94	0.28	1.113	-0.211			499		Level 3
S416Q01	40.41	0.27	1.068	0.217			573		Level 4
S421Q01	40.64	0.27	1.068	0.181			567		Level 4
S421Q02	27.11	0.25	0.402	1.524			868		Level 6
S421Q03	56.73	0.27	1.362	-0.283			479		Level 2
S425Q02	47.19	0.29	1.000	0.094			555		Level 3
S425Q03	39.20	0.27	0.915	0.206			578		Level 4
S425Q04	27.07	0.26	1.114	0.647			643		Level 5
S425Q05	62.24	0.27	1.000	-0.452			463		Level 2
S428Q01	52.53	0.27	1.334	-0.180			497		Level 3
S428Q03	67.44	0.26	1.683	-0.498			436		Level 2
S428Q05	39.23	0.27	1.756	0.108			537		Level 3



[Part 8/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	CBA international % correct	CBA international % correct S.E.	CBA item parameters (RP = 0.50)				CBA thresholds (RP = .62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
S437Q01	66.40	0.25	1.113	-0.568			439		Level 2
S437Q03	47.96	0.27	0.684	0.094			577		Level 4
S437Q04	47.96	0.27	1.000	-0.047			531		Level 3
S437Q06	69.29	0.25	1.809	-0.458			440		Level 2
S438Q01	71.99	0.24	1.138	-0.764			405		Level 1a
S438Q02	57.19	0.27	1.181	-0.291			483		Level 2
S438Q03	30.75	0.25	1.394	0.350			584		Level 4
S458Q01	13.20	0.19	1.032	1.284			754		Level 6
S458Q02	45.38	0.27	0.994	0.013			542		Level 3
S465Q01	38.68	0.23	0.985	0.178	-0.077	0.077	521	599	Level 4
S465Q02	55.57	0.27	0.922	-0.256			500		Level 3
S465Q04	36.29	0.30	0.562	0.561			671		Level 5
S466Q01	64.33	0.27	1.000	-0.578			442		Level 2
S466Q05	49.42	0.28	0.693	-0.071			549		Level 3
S466Q07	65.83	0.27	0.772	-0.752			427		Level 2
S476Q01	67.07	0.26	1.000	-0.617			435		Level 2
S476Q02	65.58	0.25	0.878	-0.729			423		Level 2
S476Q03	56.40	0.27	1.183	-0.302			481		Level 2
S478Q01	42.75	0.27	0.714	0.229			597		Level 4
S478Q02	47.25	0.27	1.200	-0.054			522		Level 3
S478Q03	60.37	0.26	0.974	-0.432			468		Level 2
S495Q01	35.71	0.25	0.903	0.409			613		Level 4
S495Q02	55.01	0.26	0.856	-0.251			505		Level 3
S495Q03	34.28	0.26	1.628	0.235			560		Level 4
S495Q04	31.91	0.25	0.890	0.424			616		Level 4
S498Q02	38.94	0.27	0.678	0.225			600		Level 4
S498Q03	38.92	0.27	0.497	0.487			670		Level 5
S498Q04	52.62	0.26	0.952	-0.175	-1.022	1.022	461	499	Level 3
S510Q01	46.62	0.27	0.587	0.085			588		Level 4
S510Q04	36.85	0.25	1.176	0.321			586		Level 4
S514Q02	72.97	0.25	1.682	-0.672			406		Level 1a
S514Q03	36.26	0.27	1.196	0.262			575		Level 4
S514Q04	49.01	0.28	1.914	-0.116			497		Level 3
S519Q01	31.42	0.22	0.742	0.383	-0.246	0.246	555	648	Level 5
S519Q02	48.36	0.27	0.492	-0.200			556		Level 3
S519Q03	23.89	0.22	0.838	0.866			694		Level 5
S521Q02	50.07	0.27	0.599	-0.121			551		Level 3
S521Q06	86.63	0.18	1.545	-1.100			337		Level 1a
S524Q06	59.15	0.26	0.977	-0.421			469		Level 2
S524Q07	32.43	0.25	1.382	0.437			599		Level 4
S527Q01	11.81	0.17	1.000	1.304			759		Level 6
S527Q03	49.25	0.27	0.634	-0.067			556		Level 3
S527Q04	48.20	0.27	0.846	-0.054			539		Level 3

[Part 9/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	PBA international % correct	PBA international % correct S.E.	PBA item parameters (RP = 0.50)				PBA thresholds (RP = .62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
S131Q02	32.46	0.39	1.425	0.075			537		Level 3
S131Q04	18.25	0.30	1.211	0.551			624		Level 4
S252Q01	32.38	0.36	0.858	0.032			553		Level 3
S252Q02	64.07	0.37	0.458	-1.055			419		Level 2
S252Q03	42.45	0.40	0.708	-0.196			526		Level 3
S256Q01	81.39	0.28	1.000	-1.412			302		Level 1b
S268Q01	51.73	0.40	1.000	-0.580			442		Level 2
S268Q02	23.84	0.41	1.571	0.335			578		Level 4
S268Q06	40.93	0.39	0.916	-0.113			525		Level 3
S269Q01	34.93	0.37	1.541	-0.151			497		Level 3
S269Q03	29.40	0.36	1.625	0.230			559		Level 4
S269Q04	22.71	0.31	0.796	0.637			659		Level 5
S304Q01	26.79	0.36	1.387	0.149			551		Level 3
S304Q02	44.38	0.41	1.000	-0.323			485		Level 3
S304Q03	26.80	0.37	1.588	0.258			565		Level 4
S304Q03	24.41	0.37	2.475	0.009			512		Level 3
S326Q01	42.31	0.38	1.208	-0.194			498		Level 3
S326Q02	34.96	0.38	1.824	-0.206			483		Level 2
S326Q03	34.83	0.40	1.421	-0.144			501		Level 3
S326Q04	15.92	0.28	1.000	0.851			682		Level 5
S327Q01	47.28	0.40	0.991	-0.294			490		Level 3
S408Q01	46.10	0.38	0.965	-0.338			484		Level 2
S408Q03	26.56	0.33	0.983	0.618			644		Level 5
S408Q04	42.82	0.36	0.653	-0.098			548		Level 3
S408Q05	28.01	0.33	0.941	0.227			580		Level 4
S413Q04	28.02	0.33	0.963	0.249			583		Level 4
S413Q05	50.68	0.40	0.908	-0.499			460		Level 2
S413Q06	25.39	0.36	1.623	0.239			561		Level 4
S415Q02	58.05	0.41	1.409	-0.672			412		Level 2
S415Q07	56.08	0.40	0.902	-0.677			431		Level 2
S415Q08	40.74	0.39	1.113	-0.211			499		Level 3
S416Q01	28.98	0.39	1.068	0.217			573		Level 4
S421Q01	33.22	0.40	1.068	0.235			576		Level 4
S421Q02	26.69	0.35	0.402	1.224			817		Level 6
S421Q03	44.52	0.40	1.362	-0.283			479		Level 2
S425Q02	37.27	0.38	1.000	0.094			555		Level 3
S425Q03	35.19	0.38	0.915	0.206			578		Level 4
S425Q04	17.69	0.35	1.114	0.647			643		Level 5
S425Q05	51.09	0.41	1.000	-0.452			463		Level 2
S428Q01	43.06	0.38	1.334	-0.255			484		Level 2
S428Q03	47.61	0.39	1.683	-0.498			436		Level 2
S428Q05	26.48	0.36	1.756	0.108			537		Level 3



[Part 10/10]

Table A.1 PISA 2015 main survey trend science item classification

Generic ID	PBA international % correct	PBA international % correct S.E.	PBA item parameters (RP = 0.50)				PBA thresholds (RP = .62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
S437Q01	56.81	0.40	1.113	-0.568			439		Level 2
S437Q03	39.27	0.36	0.684	0.094			577		Level 4
S437Q04	40.51	0.37	1.000	-0.135			516		Level 3
S437Q06	47.30	0.39	1.809	-0.458			440		Level 2
S438Q01	54.79	0.40	1.138	-0.885			384		Level 1a
S438Q02	47.73	0.40	1.181	-0.440			458		Level 2
S438Q03	21.52	0.32	1.394	0.350			584		Level 4
S458Q01	18.12	0.32	1.032	1.032			711		Level 6
S458Q02	35.25	0.38	0.994	-0.094			524		Level 3
S465Q01	22.89	0.28	0.985	0.112	-0.080	0.080	510	588	Level 4
S465Q02	41.83	0.39	0.922	-0.256			500		Level 3
S465Q04	29.14	0.35	0.562	0.561			671		Level 5
S466Q01	51.55	0.39	1.000	-0.623			434		Level 2
S466Q05	37.89	0.39	0.693	-0.164			533		Level 3
S466Q07	48.73	0.38	0.772	-0.752			427		Level 2
S476Q01	57.30	0.41	1.000	-0.617			435		Level 2
S476Q02	58.17	0.38	0.878	-0.729			423		Level 2
S476Q03	43.44	0.39	1.183	-0.302			481		Level 2
S478Q01	33.02	0.37	0.714	0.229			597		Level 4
S478Q02	34.96	0.38	1.200	-0.054			522		Level 3
S478Q03	51.11	0.40	0.974	-0.517			453		Level 2
S495Q01	25.21	0.33	0.903	0.409			613		Level 4
S495Q02	45.34	0.39	0.856	-0.251			505		Level 3
S495Q03	23.55	0.35	1.628	0.235			560		Level 4
S495Q04	27.43	0.35	0.890	0.143			569		Level 4
S498Q02	35.99	0.39	0.678	0.225			600		Level 4
S498Q03	36.74	0.39	0.497	0.487			670		Level 5
S498Q04	41.38	0.40	0.952	-0.232	-0.631	0.631	452	501	Level 3
S510Q01	48.42	0.38	0.587	-0.175			544		Level 3
S510Q04	21.04	0.31	1.176	0.321			586		Level 4
S514Q02	50.11	0.41	1.682	-0.786			387		Level 1a
S514Q03	31.04	0.34	1.196	0.086			546		Level 3
S514Q04	29.85	0.36	1.914	-0.115			497		Level 3
S519Q01	21.99	0.30	0.742	0.383	-0.246	0.246	555	648	Level 5
S519Q02	42.65	0.37	0.492	-0.200			556		Level 3
S519Q03	19.62	0.31	0.838	0.866			694		Level 5
S521Q02	42.00	0.38	0.599	-0.121			551		Level 3
S521Q06	74.22	0.33	1.545	-1.100			337		Level 1a
S524Q06	51.47	0.39	0.977	-0.355			481		Level 2
S524Q07	21.07	0.32	1.382	0.437			599		Level 4
S527Q01	11.09	0.23	1.000	1.163			735		Level 6
S527Q03	38.62	0.37	0.634	-0.067			556		Level 3
S527Q04	37.68	0.38	0.846	-0.054			539		Level 3

[Part 1/8]

Table A.2 PISA 2015 main survey new science item classification

Generic ID	Item ID in analysis output	Unit name	Mode [paper-based (PB); computer-based (CB)]	Unit type	2015 main survey cluster	Sequence in main survey	Sequence in unit (field trial)
S601Q01	CS601Q01S	Sustainable Fish Farming	CB	Standard	S12	8	1 / 4
S601Q02	CS601Q02S	Sustainable Fish Farming	CB	Standard	S12	9	2 / 4
S601Q04	CS601Q04S	Sustainable Fish Farming	CB	Standard	S12	10	4 / 4
S602Q01	CS602Q01S	Urban Heat Island Effect	CB	Standard	S07	14	1 / 4
S602Q02	CS602Q02S	Urban Heat Island Effect	CB	Standard	S07	15	2 / 4
S602Q03	DS602Q03C	Urban Heat Island Effect	CB	Standard	S07	16	3 / 4
S602Q04	CS602Q04S	Urban Heat Island Effect	CB	Standard	S07	17	4 / 4
S603Q01	CS603Q01S	Elephants and Acacia Trees	CB	Standard	S07	9	1 / 5
S603Q02	DS603Q02C	Elephants and Acacia Trees	CB	Standard	S07	10	2 / 5
S603Q03	CS603Q03S	Elephants and Acacia Trees	CB	Standard	S07	11	3 / 5
S603Q04	CS603Q04S	Elephants and Acacia Trees	CB	Standard	S07	12	4 / 5
S603Q05	CS603Q05S	Elephants and Acacia Trees	CB	Standard	S07	13	5 / 5
S604Q02	CS604Q02S	Water from Fog	CB	Standard	S10	8	2 / 4
S604Q04	DS604Q04C	Water from Fog	CB	Standard	S10	9	4 / 4
S605Q01	CS605Q01S	Geothermal Energy	CB	Standard	S08	13	1 / 4
S605Q02	CS605Q02S	Geothermal Energy	CB	Standard	S08	14	2 / 4
S605Q03	CS605Q03S	Geothermal Energy	CB	Standard	S08	15	3 / 4
S605Q04	DS605Q04C	Geothermal Energy	CB	Standard	S08	16	4 / 4
S607Q01	CS607Q01S	Birds and Caterpillars	CB	Standard	S08	1	1 / 4
S607Q02	CS607Q02S	Birds and Caterpillars	CB	Standard	S08	2	2 / 4
S607Q03	DS607Q03C	Birds and Caterpillars	CB	Standard	S08	3	3 / 4
S608Q01	CS608Q01S	Ammonoids	CB	Standard	S08	9	1 / 4
S608Q02	CS608Q02S	Ammonoids	CB	Standard	S08	10	2 / 4
S608Q03	CS608Q03S	Ammonoids	CB	Standard	S08	11	3 / 4
S608Q04	DS608Q04C	Ammonoids	CB	Standard	S08	12	4 / 4
S610Q01	DS610Q01C	Brain-Controlled Robotics	CB	Standard	S12	11	1 / 3
S610Q02	CS610Q02S	Brain-Controlled Robotics	CB	Standard	S12	12	2 / 3
S610Q04	CS610Q04S	Brain-Controlled Robotics	CB	Standard	S12	13	3 / 3
S615Q01	CS615Q01S	Understanding Tsunamis	CB	Interactive	S10	5	2 / 5
S615Q02	CS615Q02S	Understanding Tsunamis	CB	Interactive	S10	6	3 / 5
S615Q05	CS615Q05S	Understanding Tsunamis	CB	Interactive	S10	7	5 / 5
S615Q07	CS615Q07S	Understanding Tsunamis	CB	Interactive	S10	4	1 / 5
S620Q01	CS620Q01S	Tornadoes	CB	Standard	S09	10	2 / 5
S620Q02	CS620Q02S	Tornadoes	CB	Standard	S09	11	3 / 5
S620Q04	DS620Q04C	Tornadoes	CB	Standard	S09	12	4 / 5
S625Q01	DS625Q01C	Wildfires and the Fire Triangle	CB	Standard	S10	1	1 / 3
S625Q02	CS625Q02S	Wildfires and the Fire Triangle	CB	Standard	S10	2	2 / 3
S625Q03	CS625Q03S	Wildfires and the Fire Triangle	CB	Standard	S10	3	3 / 3
S626Q01	CS626Q01S	Sounds in Marine Habitats	CB	Standard	S12	14	1 / 4
S626Q02	CS626Q02S	Sounds in Marine Habitats	CB	Standard	S12	15	2 / 4
S626Q03	CS626Q03S	Sounds in Marine Habitats	CB	Standard	S12	16	3 / 4
S626Q04	DS626Q04C	Sound in Marine Habitats	CB	Standard	S12	17	4 / 4
S627Q01	CS627Q01S	Car Tyres	CB	Standard	S07	1	1 / 3
S627Q03	CS627Q03S	Car Tyres	CB	Standard	S07	2	2 / 3
S627Q04	CS627Q04S	Car Tyres	CB	Standard	S07	3	3 / 3
S629Q01	DS629Q01C	Solar Cooker	CB	Standard	S11	9	1 / 4
S629Q02	CS629Q02S	Solar Cooker	CB	Standard	S11	10	2 / 4
S629Q03	DS629Q03C	Solar Cooker	CB	Standard	S11	11	3 / 4
S629Q04	CS629Q04S	Solar Cooker	CB	Standard	S11	12	4 / 4
S634Q01	CS634Q01S	Vaccination and Spreading of Disease	CB	Interactive	S09	5	1 / 5



[Part 2/8]

Table A.2 PISA 2015 main survey new science item classification

Generic ID	Item ID in analysis output	Unit name	Mode [paper-based (PB); computer-based (CB)]	Unit type	2015 main survey cluster	Sequence in main survey	Sequence in unit (field trial)
S634Q02	CS634Q02S	Vaccination and Spreading of Disease	CB	Interactive	S09	6	2 / 5
S634Q03	DS634Q03C	Vaccination and Spreading of Disease	CB	Interactive	S09	7	3 / 5
S634Q04	CS634Q04S	Vaccination and Spreading of Disease	CB	Interactive	S09	9	5 / 5
S634Q05	DS634Q05C	Vaccination and Spreading of Disease	CB	Interactive	S09	8	4 / 5
S635Q01	CS635Q01S	Save the Fish	CB	Interactive	S07	4	1 / 6
S635Q02	CS635Q02S	Save the Fish	CB	Interactive	S07	5	2 / 6
S635Q03	DS635Q03C	Save the Fish	CB	Interactive	S07	6	3 / 6
S635Q04	CS635Q04S	Save the Fish	CB	Interactive	S07	7	4 / 6
S635Q05	DS635Q05C	Save the Fish	CB	Interactive	S07	8	5 / 6
S637Q01	DS637Q01C	Slope-Face Investigation	CB	Standard	S12	5	1 / 4
S637Q02	CS637Q02S	Slope-Face Investigation	CB	Standard	S12	6	2 / 4
S637Q05	DS637Q05C	Slope-Face Investigation	CB	Standard	S12	7	4 / 4
S638Q01	CS638Q01S	Oil Spills	CB	Standard	S09	13	1 / 5
S638Q02	CS638Q02S	Oil Spills	CB	Standard	S09	14	2 / 5
S638Q04	CS638Q04S	Oil Spills	CB	Standard	S09	15	4 / 5
S638Q05	DS638Q05C	Oil Spills	CB	Standard	S09	16	5 / 5
S641Q01	CS641Q01S	Meteoroids and Craters	CB	Standard	S12	1	1 / 3
S641Q02	CS641Q02S	Meteoroids and Craters	CB	Standard	S12	2	2 / 3
S641Q03	CS641Q03S	Meteoroids and Craters	CB	Standard	S12	3	3 / 3
S641Q04	CS641Q04S	Meteoroids and Craters	CB	Standard	S12	4	3 / 3
S643Q01	CS643Q01S	Comparing Light Bulbs	CB	Interactive	S11	5	2 / 5
S643Q02	CS643Q02S	Comparing Light Bulbs	CB	Interactive	S11	6	3 / 5
S643Q03	DS643Q03C	Comparing Light Bulbs	CB	Interactive	S11	4	1 / 5
S643Q04	CS643Q04S	Comparing Light Bulbs	CB	Interactive	S11	7	4 / 5
S643Q05	DS643Q05C	Comparing Light Bulbs	CB	Interactive	S11	8	5 / 5
S645Q01	CS645Q01S	Carbon Dioxide in Earth's Atmosphere	CB	Standard	S10	10	1 / 4
S645Q03	CS645Q03S	Carbon Dioxide in Earth's Atmosphere	CB	Standard	S10	11	2 / 4
S645Q04	DS645Q04C	Carbon Dioxide in Earth's Atmosphere	CB	Standard	S10	12	3 / 4
S645Q05	DS645Q05C	Carbon Dioxide in Earth's Atmosphere	CB	Standard	S10	13	4 / 4
S646Q01	CS646Q01S	Nanoparticles	CB	Interactive	S08	4	1 / 5
S646Q02	CS646Q02S	Nanoparticles	CB	Interactive	S08	5	2 / 5
S646Q03	CS646Q03S	Nanoparticles	CB	Interactive	S08	6	3 / 5
S646Q04	DS646Q04C	Nanoparticles	CB	Interactive	S08	7	4 / 5
S646Q05	DS646Q05C	Nanoparticles	CB	Interactive	S08	8	5 / 5
S648Q01	DS648Q01C	Habitable Zone	CB	Standard	S11	13	1 / 5
S648Q02	CS648Q02S	Habitable Zone	CB	Standard	S11	14	2 / 5
S648Q03	CS648Q03S	Habitable Zone	CB	Standard	S11	15	3 / 5
S648Q05	DS648Q05C	Habitable Zone	CB	Standard	S11	16	5 / 5
S649Q01	CS649Q01S	Weather Balloon	CB	Standard	S09	1	1 / 4
S649Q02	DS649Q02C	Weather Balloon	CB	Standard	S09	2	2 / 4
S649Q03	CS649Q03S	Weather Balloon	CB	Standard	S09	3	3 / 4
S649Q04	CS649Q04S	Weather Balloon	CB	Standard	S09	4	4 / 4
S656Q01	CS656Q01S	Bird Migration	CB	Standard	S11	1	1 / 5
S656Q02	DS656Q02C	Bird Migration	CB	Standard	S11	2	2 / 5
S656Q04	CS656Q04S	Bird Migration	CB	Standard	S11	3	3 / 5
S657Q01	CS657Q01S	Invasive Species	CB	Standard	S10	14	1 / 4
S657Q02	CS657Q02S	Invasive Species	CB	Standard	S10	15	2 / 4
S657Q03	CS657Q03S	Invasive Species	CB	Standard	S10	16	3 / 4
S657Q04	DS657Q04C	Invasive Species	CB	Standard	S10	17	4 / 4

[Part 3/8]

Table A.2 PISA 2015 main survey new science item classification

Generic ID	Item format – CBA	Context 1 (2015)	Context 2	Competency (2015)	Knowledge (2015)	System (2015)
S601Q01	Complex Multiple Choice – Computer Scored	Local/ National	Natural Resources	Explain phenomena scientifically	Content	Living
S601Q02	Simple Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Interpret data and evidence scientifically	Content	Living
S601Q04	Simple Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Physical
S602Q01	Complex Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Interpret data and evidence scientifically	Procedural	Earth and Space
S602Q02	Complex Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Earth and Space
S602Q03	Open Response – Human Coded	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Physical
S602Q04	Complex Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Interpret data and evidence scientifically	Procedural	Living
S603Q01	Simple Multiple Choice – Computer Scored	Local/ National	Natural Resources	Interpret data and evidence scientifically	Procedural	Living
S603Q02	Open Response - Human Coded	Local/ National	Natural Resources	Evaluate and design scientific enquiry	Epistemic	Living
S603Q03	Simple Multiple Choice – Computer Scored	Local/ National	Natural Resources	Explain phenomena scientifically	Procedural	Living
S603Q04	Simple Multiple Choice – Computer Scored	Local/ National	Natural Resources	Explain phenomena scientifically	Content	Living
S603Q05	Simple Multiple Choice – Computer Scored	Local/ National	Natural Resources	Evaluate and design scientific enquiry	Procedural	Living
S604Q02	Complex Multiple Choice – Computer Scored	Global	Natural Resources	Explain phenomena scientifically	Content	Physical
S604Q04	Open Response - Human Coded	Global	Natural Resources	Evaluate and design scientific enquiry	Epistemic	Physical
S605Q01	Complex Multiple Choice – Computer Scored	Local/ National	Frontiers	Explain phenomena scientifically	Content	Earth and Space
S605Q02	Complex Multiple Choice – Computer Scored	Local/ National	Natural Resources	Interpret data and evidence scientifically	Content	Earth and Space
S605Q03	Simple Multiple Choice – Computer Scored	Global	Environmental Quality	Interpret data and evidence scientifically	Procedural	Physical
S605Q04	Open Response – Human Coded	Local/ National	Natural Resources	Explain phenomena scientifically	Content	Earth and Space
S607Q01	Simple Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Living
S607Q02	Complex Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Living
S607Q03	Open Response - Human Coded	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Living
S608Q01	Complex Multiple Choice – Computer Scored	Local/ National	Frontiers	Explain phenomena scientifically	Content	Earth and Space
S608Q02	Complex Multiple Choice – Computer Scored	Global	Frontiers	Interpret data and evidence scientifically	Epistemic	Living
S608Q03	Simple Multiple Choice – Computer Scored	Global	Frontiers	Explain phenomena scientifically	Content	Physical
S608Q04	Open Response – Human Coded	Local/ National	Natural Resources	Interpret data and evidence scientifically	Procedural	Earth and Space
S610Q01	Open Response – Human Coded	Personal	Health & Disease	Explain phenomena scientifically	Content	Living
S610Q02	Simple Multiple Choice – Computer Scored	Personal	Health & Disease	Explain phenomena scientifically	Content	Living
S610Q04	Complex Multiple Choice – Computer Scored	Personal	Frontiers	Evaluate and design scientific enquiry	Content	Living
S615Q01	Complex Multiple Choice – Computer Scored	Global	Hazards	Interpret data and evidence scientifically	Procedural	Earth and Space
S615Q02	Open Response – Computer Scored	Global	Hazards	Interpret data and evidence scientifically	Procedural	Earth and Space
S615Q05	Complex Multiple Choice – Computer Scored	Global	Hazards	Explain phenomena scientifically	Epistemic	Earth and Space
S615Q07	Complex Multiple Choice – Computer Scored	Global	Hazards	Interpret data and evidence scientifically	Procedural	Earth and Space
S620Q01	Simple Multiple Choice – Computer Scored	Local/ National	Hazards	Interpret data and evidence scientifically	Procedural	Earth and Space
S620Q02	Complex Multiple Choice – Computer Scored	Local/ National	Hazards	Interpret data and evidence scientifically	Procedural	Earth and Space
S620Q04	Open Response – Human Coded	Local/ National	Hazards	Evaluate and design scientific enquiry	Epistemic	Earth and Space
S625Q01	Open Response – Human Coded	Local/ National	Hazards	Explain phenomena scientifically	Content	Physical
S625Q02	Simple Multiple Choice – Computer Scored	Local/ National	Hazards	Explain phenomena scientifically	Content	Physical
S625Q03	Complex Multiple Choice – Computer Scored	Local/ National	Hazards	Explain phenomena scientifically	Content	Physical
S626Q01	Simple Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Physical
S626Q02	Simple Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Evaluate and design scientific enquiry	Procedural	Physical
S626Q03	Simple Multiple Choice – Computer Scored	Local/ National	Environmental Quality	Interpret data and evidence scientifically	Procedural	Physical
S626Q04	Open Response - Human Coded	Local/ National	Environmental Quality	Interpret data and evidence scientifically	Procedural	Living
S627Q01	Simple Multiple Choice - Computer Scored	Personal	Frontiers	Explain phenomena scientifically	Content	Physical
S627Q03	Complex Multiple Choice - Computer Scored	Personal	Frontiers	Explain phenomena scientifically	Content	Physical
S627Q04	Complex Multiple Choice - Computer Scored	Personal	Hazards	Evaluate and design scientific enquiry	Epistemic	Physical
S629Q01	Open Response - Human Coded	Local/ National	Natural Resources	Explain phenomena scientifically	Content	Physical
S629Q02	Complex Multiple Choice - Computer Scored	Local/ National	Natural Resources	Explain phenomena scientifically	Content	Physical
S629Q03	Open Response - Human Coded	Local/ National	Natural Resources	Interpret data and evidence scientifically	Procedural	Earth and Space
S629Q04	Complex Multiple Choice - Computer Scored	Local/ National	Natural Resources	Evaluate and design scientific enquiry	Epistemic	Physical
S634Q01	Complex Multiple Choice - Computer Scored	Global	Health & Disease	Interpret data and evidence scientifically	Procedural	Living



[Part 4/8]

Table A.2 PISA 2015 main survey new science item classification

Generic ID	Item format – CBA	Context 1 (2015)	Context 2	Competency (2015)	Knowledge (2015)	System (2015)
S634Q02	Complex Multiple Choice - Computer Scored	Global	Health & Disease	Interpret data and evidence scientifically	Procedural	Living
S634Q03	Open Response - Human Coded	Global	Health & Disease	Explain phenomena scientifically	Procedural	Living
S634Q04	Complex Multiple Choice - Computer Scored	Global	Health & Disease	Evaluate and design scientific enquiry	Epistemic	Living
S634Q05	Open Response - Human Coded	Global	Health & Disease	Evaluate and design scientific enquiry	Epistemic	Living
S635Q01	Complex Multiple Choice - Computer Scored	Local/ National	Natural Resources	Explain phenomena scientifically	Content	Living
S635Q02	Complex Multiple Choice - Computer Scored	Local/ National	Natural Resources	Evaluate and design scientific enquiry	Procedural	Living
S635Q03	Open Response - Human Coded	Local/ National	Natural Resources	Evaluate and design scientific enquiry	Procedural	Living
S635Q04	Open Response - Computer Scored	Local/ National	Natural Resources	Evaluate and design scientific enquiry	Procedural	Living
S635Q05	Open Response - Human Coded	Local/ National	Natural Resources	Explain phenomena scientifically	Procedural	Living
S637Q01	Open Response - Human Coded	Local/ National	Natural Resources	Evaluate and design scientific enquiry	Epistemic	Earth and Space
S637Q02	Complex Multiple Choice - Computer Scored	Local/ National	Natural Resources	Evaluate and design scientific enquiry	Epistemic	Earth and Space
S637Q05	Open Response - Human Coded	Local/ National	Natural Resources	Interpret data and evidence scientifically	Epistemic	Earth and Space
S638Q01	Complex Multiple Choice - Computer Scored	Global	Environmental Quality	Explain phenomena scientifically	Content	Earth and Space
S638Q02	Complex Multiple Choice - Computer Scored	Global	Environmental Quality	Explain phenomena scientifically	Content	Earth and Space
S638Q04	Complex Multiple Choice - Computer Scored	Global	Frontiers	Evaluate and design scientific enquiry	Epistemic	Living
S638Q05	Open Response - Human Coded	Global	Frontiers	Explain phenomena scientifically	Content	Living
S641Q01	Simple Multiple Choice - Computer Scored	Global	Frontiers	Explain phenomena scientifically	Content	Physical
S641Q02	Complex Multiple Choice - Computer Scored	Global	Frontiers	Explain phenomena scientifically	Content	Earth and Space
S641Q03	Complex Multiple Choice - Computer Scored	Global	Frontiers	Interpret data and evidence scientifically	Content	Earth and Space
S641Q04	Complex Multiple Choice - Computer Scored	Global	Frontiers	Interpret data and evidence scientifically	Content	Earth and Space
S643Q01	Simple Multiple Choice - Computer Scored	Personal	Frontiers	Interpret data and evidence scientifically	Procedural	Physical
S643Q02	Complex Multiple Choice - Computer Scored	Personal	Frontiers	Evaluate and design scientific enquiry	Procedural	Physical
S643Q03	Open Response - Human Coded	Personal	Frontiers	Explain phenomena scientifically	Content	Physical
S643Q04	Complex Multiple Choice - Computer Scored	Personal	Frontiers	Evaluate and design scientific enquiry	Procedural	Physical
S643Q05	Open Response - Human Coded	Personal	Frontiers	Evaluate and design scientific enquiry	Epistemic	Physical
S645Q01	Complex Multiple Choice - Computer Scored	Global	Natural Resources	Explain phenomena scientifically	Content	Earth and Space
S645Q03	Complex Multiple Choice - Computer Scored	Global	Natural Resources	Explain phenomena scientifically	Content	Earth and Space
S645Q04	Open Response - Human Coded	Global	Natural Resources	Explain phenomena scientifically	Content	Earth and Space
S645Q05	Open Response - Human Coded	Global	Natural Resources	Explain phenomena scientifically	Content	Earth and Space
S646Q01	Simple Multiple Choice - Computer Scored	Global	Frontiers	Interpret data and evidence scientifically	Procedural	Physical
S646Q02	Open Response - Computer Scored	Global	Frontiers	Evaluate and design scientific enquiry	Procedural	Physical
S646Q03	Simple Multiple Choice - Computer Scored	Global	Frontiers	Evaluate and design scientific enquiry	Procedural	Physical
S646Q04	Open Response - Human Coded	Global	Frontiers	Explain phenomena scientifically	Procedural	Physical
S646Q05	Open Response - Human Coded	Global	Frontiers	Evaluate and design scientific enquiry	Epistemic	Physical
S648Q01	Open Response - Human Coded	Global	Frontiers	Interpret data and evidence scientifically	Procedural	Earth and Space
S648Q02	Complex Multiple Choice - Computer Scored	Global	Frontiers	Interpret data and evidence scientifically	Procedural	Earth and Space
S648Q03	Complex Multiple Choice - Computer Scored	Global	Frontiers	Interpret data and evidence scientifically	Procedural	Earth and Space
S648Q05	Open Response - Human Coded	Global	Frontiers	Explain phenomena scientifically	Epistemic	Earth and Space
S649Q01	Simple Multiple Choice - Computer Scored	Local/ National	Frontiers	Explain phenomena scientifically	Content	Physical
S649Q02	Open Response - Human Coded	Local/ National	Frontiers	Explain phenomena scientifically	Content	Physical
S649Q03	Complex Multiple Choice - Computer Scored	Local/ National	Frontiers	Explain phenomena scientifically	Content	Physical
S649Q04	Simple Multiple Choice - Computer Scored	Local/ National	Frontiers	Explain phenomena scientifically	Content	Physical
S656Q01	Simple Multiple Choice - Computer Scored	Global	Environmental Quality	Explain phenomena scientifically	Content	Living
S656Q02	Open Response - Human Coded	Global	Environmental Quality	Evaluate and design scientific enquiry	Procedural	Living
S656Q04	Complex Multiple Choice - Computer Scored	Global	Environmental Quality	Interpret data and evidence scientifically	Procedural	Living
S657Q01	Complex Multiple Choice - Computer Scored	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Living
S657Q02	Simple Multiple Choice - Computer Scored	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Living
S657Q03	Simple Multiple Choice - Computer Scored	Local/ National	Environmental Quality	Interpret data and evidence scientifically	Procedural	Living
S657Q04	Open Response - Human Coded	Local/ National	Environmental Quality	Explain phenomena scientifically	Content	Living

[Part 5/8]

Table A.2 PISA 2015 main survey new science item classification

Generic ID	Depth of knowledge	Source	Unit origin	Language of submission	International % correct	International % correct S.E.
S601Q01	Medium	2015	Netherlands	English	5.48	0.09
S601Q02	Low	2015	Netherlands	English	62.79	0.20
S601Q04	Low	2015	Netherlands	English	35.48	0.19
S602Q01	Low	2015	University of Luxembourg (International Test Development Team)	English	76.40	0.17
S602Q02	Low	2015	University of Luxembourg (International Test Development Team)	English	30.24	0.19
S602Q03	Medium	2015	University of Luxembourg (International Test Development Team)	English	23.55	0.17
S602Q04	Low	2015	University of Luxembourg (International Test Development Team)	English	67.53	0.20
S603Q01	Low	2015	Singapore	English	69.68	0.19
S603Q02	Medium	2015	Singapore	English	31.29	0.19
S603Q03	Low	2015	Singapore	English	66.03	0.19
S603Q04	Medium	2015	Singapore	English	52.66	0.20
S603Q05	Medium	2015	Singapore	English	52.66	0.20
S604Q02	Medium	2015	Singapore	English	40.70	0.19
S604Q04	Medium	2015	Singapore	English	24.65	0.17
S605Q01	Medium	2015	France	French	37.80	0.19
S605Q02	Medium	2015	France	French	28.75	0.19
S605Q03	Medium	2015	France	French	51.94	0.20
S605Q04	Medium	2015	France	French	52.23	0.20
S607Q01	Low	2015	Singapore	English	79.42	0.16
S607Q02	Low	2015	Singapore	English	43.81	0.20
S607Q03	High	2015	Singapore	English	38.35	0.17
S608Q01	Low	2015	France	French	32.48	0.18
S608Q02	Medium	2015	France	French	56.40	0.20
S608Q03	Medium	2015	France	French	40.08	0.20
S608Q04	Medium	2015	France	French	42.36	0.21
S610Q01	Medium	2015	Spain	Spanish	25.93	0.17
S610Q02	Low	2015	Spain	Spanish	82.10	0.15
S610Q04	Medium	2015	Spain	Spanish	43.85	0.20
S615Q01	Medium	2015	Chinese Taipei	English	76.93	0.17
S615Q02	Medium	2015	Chinese Taipei	English	38.33	0.20
S615Q05	Medium	2015	Chinese Taipei	English	17.36	0.15
S615Q07	Medium	2015	Chinese Taipei	English	26.86	0.18
S620Q01	Medium	2015	Czech Republic	English	78.08	0.17
S620Q02	Medium	2015	Czech Republic	English	32.86	0.19
S620Q04	High	2015	Czech Republic	English	31.09	0.19
S625Q01	Low	2015	Australia	English	39.57	0.20
S625Q02	Low	2015	Australia	English	56.76	0.19
S625Q03	Medium	2015	Australia	English	51.44	0.20
S626Q01	Medium	2015	University of Luxembourg (International Test Development Team)	English	56.15	0.20
S626Q02	Medium	2015	University of Luxembourg (International Test Development Team)	English	48.27	0.20
S626Q03	Medium	2015	University of Luxembourg (International Test Development Team)	English	63.51	0.20
S626Q04	Medium	2015	University of Luxembourg (International Test Development Team)	English	48.34	0.21
S627Q01	Low	2015	Korea	English	39.80	0.19
S627Q03	Medium	2015	Korea	English	71.32	0.18
S627Q04	Low	2015	Korea	English	57.69	0.19
S629Q01	Medium	2015	Viet Nam	English	54.31	0.18
S629Q02	Medium	2015	Viet Nam	English	37.96	0.19
S629Q03	Medium	2015	Viet Nam	English	49.72	0.20
S629Q04	Medium	2015	Viet Nam	English	50.35	0.20
S634Q01	Medium	2015	Israel - CET (International Test Development Team)	English	15.96	0.15



[Part 6/8]

Table A.2 PISA 2015 main survey new science item classification

Generic ID	Depth of knowledge	Source	Unit origin	Language of submission	International % correct	International % correct S.E.
S634Q02	Medium	2015	Israel - CET (International Test Development Team)	English	30.60	0.15
S634Q03	Medium	2015	Israel - CET (International Test Development Team)	English	16.49	0.15
S634Q04	High	2015	Israel - CET (International Test Development Team)	English	44.30	0.20
S634Q05	Medium	2015	Israel - CET (International Test Development Team)	English	10.27	0.12
S635Q01	Low	2015	Australia	English	51.32	0.17
S635Q02	Low	2015	Australia	English	66.07	0.19
S635Q03	Medium	2015	Australia	English	36.14	0.19
S635Q04	High	2015	Australia	English	41.46	0.16
S635Q05	High	2015	Australia	English	13.74	0.13
S637Q01	Medium	2015	Israel - CET (International Test Development Team)	English	47.51	0.20
S637Q02	Medium	2015	Israel - CET (International Test Development Team)	English	14.53	0.10
S637Q05	High	2015	Israel - CET (International Test Development Team)	English	31.89	0.19
S638Q01	Medium	2015	Korea	English	47.75	0.20
S638Q02	Low	2015	Korea	English	72.44	0.18
S638Q04	Medium	2015	Korea	English	26.98	0.18
S638Q05	Medium	2015	Korea	English	49.56	0.26
S641Q01	Low	2015	US - ETS (International Test Development Team)	English	55.83	0.19
S641Q02	Low	2015	US - ETS (International Test Development Team)	English	64.76	0.19
S641Q03	Low	2015	US - ETS (International Test Development Team)	English	88.29	0.13
S641Q04	Medium	2015	US - ETS (International Test Development Team)	English	67.05	0.19
S643Q01	Medium	2015	Japan	Japanese	64.15	0.19
S643Q02	Medium	2015	Japan	Japanese	50.59	0.20
S643Q03	Low	2015	Japan	Japanese	28.55	0.19
S643Q04	Medium	2015	Japan	Japanese	24.47	0.17
S643Q05	Medium	2015	Japan	Japanese	21.06	0.16
S645Q01	Low	2015	Spain	Spanish	50.05	0.18
S645Q03	Medium	2015	Spain	Spanish	51.84	0.20
S645Q04	Medium	2015	Spain	Spanish	49.51	0.20
S645Q05	Medium	2015	Spain	Spanish	22.39	0.16
S646Q01	Low	2015	Chinese Taipei	English	75.69	0.17
S646Q02	Medium	2015	Chinese Taipei	English	47.16	0.20
S646Q03	Medium	2015	Chinese Taipei	English	65.91	0.19
S646Q04	High	2015	Chinese Taipei	English	25.30	0.18
S646Q05	Medium	2015	Chinese Taipei	English	11.48	0.16
S648Q01	Medium	2015	US - ETS (International Test Development Team)	English	33.66	0.20
S648Q02	Medium	2015	US - ETS (International Test Development Team)	English	37.24	0.19
S648Q03	Medium	2015	US - ETS (International Test Development Team)	English	57.29	0.20
S648Q05	Medium	2015	US - ETS (International Test Development Team)	English	35.53	0.22
S649Q01	Medium	2015	Sweden	English	24.89	0.17
S649Q02	Low	2015	Sweden	English	14.51	0.14
S649Q03	Medium	2015	Sweden	English	27.21	0.18
S649Q04	Medium	2015	Sweden	English	42.52	0.20
S656Q01	Medium	2015	Netherlands	English	54.14	0.20
S656Q02	High	2015	Netherlands	English	27.69	0.18
S656Q04	Medium	2015	Netherlands	English	36.25	0.19
S657Q01	Low	2015	University of Luxembourg (International Test Development Team)	English	67.14	0.19
S657Q02	Medium	2015	University of Luxembourg (International Test Development Team)	English	37.32	0.20
S657Q03	Medium	2015	University of Luxembourg (International Test Development Team)	English	45.36	0.21
S657Q04	Medium	2015	University of Luxembourg (International Test Development Team)	English	26.63	0.17

## [Part 7/8]

Table A.2 PISA 2015 main survey new science item classification

Generic ID	Item parameters (RP = 0.50)				Thresholds (RP = 0.62)		Level
	Slope	Difficulty	Step 1	Step 2	1,00	2,00	
S601Q01	1.652	1.306			740		Level 6
S601Q02	1.375	-0.413			456		Level 2
S601Q04	1.033	0.283			585		Level 4
S602Q01	1.440	-0.808			388		Level 1a
S602Q02	0.810	0.613			654		Level 5
S602Q03	1.061	0.714			657		Level 5
S602Q04	1.277	-0.558			435		Level 2
S603Q01	1.759	-0.540			427		Level 2
S603Q02	1.518	0.328			578		Level 4
S603Q03	1.114	-0.528			445		Level 2
S603Q04	0.721	-0.201			524		Level 3
S603Q05	1.139	-0.172			504		Level 3
S604Q02	0.979	0.172			569		Level 4
S604Q04	1.232	0.590			629		Level 4
S605Q01	1.040	0.231			576		Level 4
S605Q02	1.578	0.388			587		Level 4
S605Q03	0.683	-0.182			531		Level 3
S605Q04	1.157	-0.096	0.027	-0.027	475	550	Level 3
S607Q01	1.290	-0.931			372		Level 1a
S607Q02	1.646	0.045			528		Level 3
S607Q03	0.814	0.198	-0.308	0.308	524	603	Level 4
S608Q01	0.577	0.794			708		Level 6
S608Q02	1.381	-0.270			480		Level 2
S608Q03	0.671	0.265			608		Level 4
S608Q04	1.901	0.047			524		Level 3
S610Q01	1.269	0.577			626		Level 4
S610Q02	1.266	-1.039			354		Level 1a
S610Q04	1.581	0.037			528		Level 3
S615Q01	1.461	-0.805			388		Level 1a
S615Q02	1.604	0.146			546		Level 3
S615Q05	0.478	1.953			921		Level 6
S615Q07	1.497	0.450			599		Level 4
S620Q01	1.058	-0.967			374		Level 1a
S620Q02	0.994	0.414			609		Level 4
S620Q04	1.245	0.397			596		Level 4
S625Q01	1.406	0.140			549		Level 3
S625Q02	0.830	-0.431			477		Level 2
S625Q03	1.069	-0.186			505		Level 3
S626Q01	0.796	-0.334			495		Level 3
S626Q02	0.933	-0.056			533		Level 3
S626Q03	1.142	-0.466			455		Level 2
S626Q04	2.053	-0.066			503		Level 3
S627Q01	0.672	0.376			626		Level 4
S627Q03	0.932	-0.798			408		Level 1a
S627Q04	0.907	-0.400			477		Level 2
S629Q01	0.700	-0.198	-0.392	0.392	457	547	Level 3
S629Q02	0.918	0.248			585		Level 4
S629Q03	1.396	-0.111			507		Level 3
S629Q04	0.857	-0.137			524		Level 3
S634Q01	1.581	0.793			655		Level 5



[Part 8/8]

Table A.2 PISA 2015 main survey new science item classification

Generic ID	Item parameters ( $RP = 0.50$ )				Thresholds ( $RP = 0.62$ )		Level
	Slope	Difficulty	Step 1	Step 2	1.00	2.00	
S634Q02	1.152	0.431	0.301	-0.301	563	669	Level 5
S634Q03	1.630	0.762			649		Level 5
S634Q04	1.497	0.024			527		Level 3
S634Q05	1.547	1.050			699		Level 5
S635Q01	0.678	-0.166	0.103	-0.103	463	596	Level 4
S635Q02	1.365	-0.504			441		Level 2
S635Q03	2.188	0.166			541		Level 3
S635Q04	1.061	0.111	0.271	-0.271	509	617	Level 4
S635Q05	1.487	0.752	-0.212	0.212	617	658	Level 5
S637Q01	1.356	-0.054			517		Level 3
S637Q02	0.486	1.955	0.512	-0.512	820	1045	Level 6
S637Q05	1.334	0.370			589		Level 4
S638Q01	1.530	-0.051			514		Level 3
S638Q02	1.405	-0.667			413		Level 2
S638Q04	1.069	0.592			636		Level 5
S638Q05	1.571	0.048			530		Level 3
S641Q01	0.649	-0.490			483		Level 2
S641Q02	1.036	-0.520			450		Level 2
S641Q03	1.133	-1.395			299		Level 1b
S641Q04	1.316	-0.530			438		Level 2
S643Q01	2.177	-0.392			447		Level 2
S643Q02	1.993	-0.111			496		Level 3
S643Q03	1.427	0.476			605		Level 4
S643Q04	1.809	0.465			596		Level 4
S643Q05	1.086	0.794			669		Level 5
S645Q01	0.742	-0.127	-0.329	0.329	469	557	Level 3
S645Q03	1.092	-0.201			501		Level 3
S645Q04	1.722	-0.084			505		Level 3
S645Q05	1.287	0.847			671		Level 5
S646Q01	1.892	-0.692			400		Level 1a
S646Q02	1.243	-0.045			522		Level 3
S646Q03	1.335	-0.511			441		Level 2
S646Q04	1.513	0.523			611		Level 4
S646Q05	1.549	0.997			690		Level 5
S648Q01	1.898	0.247			558		Level 3
S648Q02	0.946	0.277			589		Level 4
S648Q03	0.511	-0.472			506		Level 3
S648Q05	1.445	0.369			586		Level 4
S649Q01	0.516	1.255			796		Level 6
S649Q02	1.061	1.279			752		Level 6
S649Q03	1.151	0.541			624		Level 4
S649Q04	0.630	0.222			605		Level 4
S656Q01	1.004	-0.228			501		Level 3
S656Q02	1.046	0.552			630		Level 4
S656Q04	1.175	0.247			574		Level 4
S657Q01	0.756	-0.814			418		Level 2
S657Q02	0.686	0.423			633		Level 4
S657Q03	1.107	0.031			540		Level 3
S657Q04	0.959	0.640	0.150	-0.150	598	701	Level 5

## [Part 1/8]

Table A.3 PISA 2015 main survey trend reading item classification

Generic ID	CBA item ID in main survey analysis output	PBA item ID in main survey analysis output	Unit name	Mode [paper-based (PB); computer-based (CB)]	2015 field trial and main survey cluster	Sequence in cluster	Sequence in unit	Item format – CBA
R055Q01	CR055Q01S	PR055Q01S	Drugged Spiders	PB and CB	R02	9	1/4	Simple Multiple Choice – Computer Scored
R055Q02	DR055Q02C	R055Q02	Drugged Spiders	PB and CB	R02	10	2/4	Open Response – Human Coded
R055Q03	DR055Q03C	R055Q03	Drugged Spiders	PB and CB	R02	11	3/4	Open Response – Human Coded
R055Q05	DR055Q05C	R055Q05	Drugged Spiders	PB and CB	R02	12	4/4	Open Response – Human Coded
R067Q01	CR067Q01S	PR067Q01S	Aesop	PB and CB	R01	8	1/3	Simple Multiple Choice – Computer Scored
R067Q04	DR067Q04C	R067Q04	Aesop	PB and CB	R01	9	2/3	Open Response – Human Coded
R067Q05	DR067Q05C	R067Q05	Aesop	PB and CB	R01	10	3/3	Open Response – Human Coded
R083Q01	CR083Q01S	PR083Q01S	Household Work	PB and CB	R06A	1	1/4	Simple Multiple Choice – Computer Scored
R083Q02	CR083Q02S	R083Q02	Household Work	PB and CB	R06A	2	2/4	Open Response – Computer Scored
R083Q03	CR083Q03S	R083Q03	Household Work	PB and CB	R06A	3	3/4	Simple Multiple Choice – Computer Scored
R083Q04	CR083Q04S	PR083Q04S	Household Work	PB and CB	R06A	4	4/4	Simple Multiple Choice – Computer Scored
R101Q01	CR101Q01S	PR101Q01S	Rhino	PB and CB	R06A	12	1/5	Simple Multiple Choice – Computer Scored
R101Q02	CR101Q02S	PR101Q02S	Rhino	PB and CB	R06A	13	2/5	Simple Multiple Choice – Computer Scored
R101Q03	CR101Q03S	PR101Q03S	Rhino	PB and CB	R06A	14	3/5	Simple Multiple Choice – Computer Scored
R101Q04	CR101Q04S	PR101Q04S	Rhino	PB and CB	R06A	15	4/5	Simple Multiple Choice – Computer Scored
R101Q05	CR101Q05S	PR101Q05S	Rhino	PB and CB	R06A	16	5/5	Simple Multiple Choice – Computer Scored
R102Q04	DR102Q04C	R102Q04A	Shirts	PB and CB	R01	7	1/3	Open Response – Human Coded
R102Q05	DR102Q05C	R102Q05	Shirts	PB and CB	R01	8	2/3	Open Response – Human Coded
R102Q07	CR102Q07S	PR102Q07S	Shirts	PB and CB	R01	9	3/3	Simple Multiple Choice – Computer Scored
R104Q01	CR104Q01S	R104Q01	Telephone	PB and CB	R02	13	1/3	Open Response – Computer Scored
R104Q02	CR104Q02S	R104Q02	Telephone	PB and CB	R02	14	2/3	Open Response – Computer Scored
R104Q05	CR104Q05S	R104Q05	Telephone	PB and CB	R02	15	3/3	Open Response – Computer Scored
R111Q01	CR111Q01S	PR111Q01S	Exchange	PB and CB	R02	5	1/3	Simple Multiple Choice – Computer Scored
R111Q02	DR111Q02BC	R111Q02B	Exchange	PB and CB	R02	6	2/3	Open Response – Human Coded
R111Q06	DR111Q06C	R111Q06B	Exchange	PB and CB	R02	8	3/3	Open Response – Human Coded
R219Q01	DR219Q01C	PR219Q01S	Employment	PB and CB	R01	1	1/2	Open Response – Human Coded
R219Q01	DR219Q01EC	R219Q01E	Employment	PB and CB	R01	2	1/2	Open Response – Human Coded
R219Q02	DR219Q02C	R219Q02	Employment	PB and CB	R01	3	2/2	Open Response – Human Coded
R220Q01	CR220Q01S	R220Q01	South Pole	PB and CB	R01	10	1/5	Open Response – Computer Scored
R220Q02	CR220Q02S	PR220Q02BS	South Pole	PB and CB	R01	11	2/5	Simple Multiple Choice – Computer Scored
R220Q04	CR220Q04S	PR220Q04S	South Pole	PB and CB	R01	12	3/5	Simple Multiple Choice – Computer Scored
R220Q05	CR220Q05S	PR220Q05S	South Pole	PB and CB	R01	13	4/5	Simple Multiple Choice – Computer Scored
R220Q06	CR220Q06S	PR220Q06S	South Pole	PB and CB	R01	14	5/5	Simple Multiple Choice – Computer Scored
R227Q01	CR227Q01S	PR227Q01S	Optician	PB and CB	R02	1	1/4	Simple Multiple Choice – Computer Scored
R227Q02	CR227Q02S	PR227Q02S	Optician	PB and CB	R02	2	2/4	Complex Multiple Choice – Computer Scored
R227Q03	DR227Q03C	R227Q03	Optician	PB and CB	R02	3	3/4	Open Response – Human Coded
R227Q06	DR227Q06C	R227Q06	Optician	PB and CB	R02	4	4/4	Open Response – Human Coded
R245Q01	CR245Q01S	R245Q01	Movie Reviews	PB and CB	R06A	10	1/2	Complex Multiple Choice – Computer Scored
R245Q02	CR245Q02S	R245Q02	Movie Reviews	PB and CB	R06A	11	2/2	Complex Multiple Choice – Computer Scored
R404Q03	CR404Q03S	PR404Q03S	Sleep	PB and CB	R05	4	1/4	Simple Multiple Choice – Computer Scored
R404Q06	CR404Q06S	PR404Q06S	Sleep	PB and CB	R05	5	2/4	Simple Multiple Choice – Computer Scored
R404Q07	CR404Q07S	PR404Q07S	Sleep	PB and CB	R05	6	3/4	Complex Multiple Choice – Computer Scored
R404Q10	DR404Q10AC	PR404Q10A	Sleep	PB and CB	R05	7	4/4	Open Response – Human Coded
R404Q10	DR404Q10BC	PR404Q10B	Sleep	PB and CB	R05	8	4/4	Open Response – Human Coded
R406Q01	DR406Q01C	PR406Q01	Kokeshi Dolls	PB and CB	R05	10	1/3	Open Response – Human Coded
R406Q02	DR406Q02C	PR406Q02	Kokeshi Dolls	PB and CB	R05	12	3/3	Open Response – Human Coded
R406Q05	DR406Q05C	PR406Q05	Kokeshi Dolls	PB and CB	R05	11	2/3	Open Response – Human Coded
R412Q01	CR412Q01S	PR412Q01S	World Languages	PB and CB	R03	9	1/4	Simple Multiple Choice – Computer Scored
R412Q05	CR412Q05S	PR412Q05S	World Languages	PB and CB	R03	10	2/4	Simple Multiple Choice – Computer Scored
R412Q06	CR412Q06S	PR412Q06S	World Languages	PB and CB	R03	12	4/4	Complex Multiple Choice – Computer Scored
R412Q08	DR412Q08C	PR412Q08	World Languages	PB and CB	R03	11	3/4	Open Response – Human Coded
R420Q02	DR420Q02C	PR420Q02	Children's Futures	PB and CB	R03	1	1/4	Open Response – Human Coded



[Part 2/8]

Table A.3 PISA 2015 main survey trend reading item classification

Generic ID	CBA item ID in main survey analysis output	PBA item ID in main survey analysis output	Unit name	Mode [paper-based (PB); computer-based (CB)]	2015 field trial and main survey cluster	Sequence in cluster	Sequence in unit	Item format – CBA
R420Q06	DR420Q06C	PR420Q06	Children's Futures	PB and CB	R03	3	3/4	Open Response – Human Coded
R420Q09	DR420Q09C	PR420Q09	Children's Futures	PB and CB	R03	4	4/4	Open Response – Human Coded
R420Q10	DR420Q10C	PR420Q10	Children's Futures	PB and CB	R03	2	2/4	Open Response – Human Coded
R424Q02	CR424Q02S	PR424Q02S	Fair Trade	PB and CB	R05	1	1/3	Complex Multiple Choice – Computer Scored
R424Q03	CR424Q03S	PR424Q03S	Fair Trade	PB and CB	R05	2	2/3	Simple Multiple Choice – Computer Scored
R424Q07	CR424Q07S	PR424Q07S	Fair Trade	PB and CB	R05	3	3/3	Simple Multiple Choice – Computer Scored
R432Q01	DR432Q01C	PR432Q01	About a book	PB and CB	R04	9	1/3	Open Response – Human Coded
R432Q05	DR432Q05C	PR432Q05	About a book	PB and CB	R04	10	2/3	Open Response – Human Coded
R432Q06	CR432Q06S	PR432Q06S	About a book	PB and CB	R04	11	3/3	Complex Multiple Choice – Computer Scored
R435Q01	CR435Q01S	PR435Q01S	Dust Mites	PB and CB	R06B	9	2/4	Simple Multiple Choice – Computer Scored
R435Q02	CR435Q02S	R435Q02	Dust Mites	PB and CB	R06B	8	1/4	Open Response – Computer Scored
R435Q05	DR435Q05C	R435Q05	Dust Mites	PB and CB	R06B	10	3/4	Open Response – Human Coded
R435Q08	CR435Q08S	PR435Q08S	Dust Mites	PB and CB	R06B	11	4/4	Complex Multiple Choice – Computer Scored
R437Q01	CR437Q01S	PR437Q01S	Narcissus	PB and CB	R03	13	1/3	Simple Multiple Choice – Computer Scored
R437Q06	CR437Q06S	PR437Q06S	Narcissus	PB and CB	R03	15	3/3	Simple Multiple Choice – Computer Scored
R437Q07	DR437Q07C	PR437Q07	Narcissus	PB and CB	R03	14	2/3	Open Response – Human Coded
R442Q02	DR442Q02C	R442Q02	Galileo	PB and CB	R06A	5	1/5	Open Response – Human Coded
R442Q03	DR442Q03C	R442Q03	Galileo	PB and CB	R06A	6	2/5	Open Response – Human Coded
R442Q05	DR442Q05C	R442Q05	Galileo	PB and CB	R06A	7	3/5	Open Response – Human Coded
R442Q06	DR442Q06C	R442Q06	Galileo	PB and CB	R06A	8	4/5	Open Response – Human Coded
R442Q07	CR442Q07S	PR442Q07S	Galileo	PB and CB	R06A	9	5/5	Simple Multiple Choice – Computer Scored
R445Q01	DR445Q01C	R445Q01	Road	PB and CB	R06B	2	2/4	Open Response – Human Coded
R445Q03	CR445Q03S	PR445Q03S	Road	PB and CB	R06B	1	1/4	Simple Multiple Choice – Computer Scored
R445Q04	CR445Q04S	PR445Q04S	Road	PB and CB	R06B	3	3/4	Simple Multiple Choice – Computer Scored
R445Q06	CR445Q06S	PR445Q06S	Road	PB and CB	R06B	4	4/4	Simple Multiple Choice – Computer Scored
R446Q03	CR446Q03S	PR446Q03	Job Vacancy	PB and CB	R04	7	1/2	Complex Multiple Choice – Computer Scored
R446Q06	DR446Q06C	PR446Q06	Job Vacancy	PB and CB	R04	8	2/2	Open Response – Human Coded
R453Q01	CR453Q01S	PR453Q01S	Summer Job	PB and CB	R03	5	1/4	Simple Multiple Choice – Computer Scored
R453Q04	DR453Q04C	PR453Q04	Summer Job	PB and CB	R03	6	2/4	Open Response – Human Coded
R453Q05	CR453Q05S	PR453Q05S	Summer Job	PB and CB	R03	7	3/4	Complex Multiple Choice – Computer Scored
R453Q06	DR453Q06C	PR453Q06	Summer Job	PB and CB	R03	8	4/4	Open Response – Human Coded
R455Q02	DR455Q02C	PR455Q02	Chocolate and Health	PB and CB	R05, U3	13	1/4	Open Response – Human Coded
R455Q03	DR455Q03C	PR455Q03	Chocolate and Health	PB and CB	R05, U3	14	2/4	Open Response – Human Coded
R455Q04	CR455Q04S	PR455Q04S	Chocolate and Health	PB and CB	R05, U3	15	3/4	Simple Multiple Choice – Computer Scored
R455Q05	CR455Q05S	PR455Q05S	Chocolate and Health	PB and CB	R05, U3	16	4/4	Complex Multiple Choice – Computer Scored
R456Q01	CR456Q01S	PR456Q01S	Biscuits	PB and CB	R04, U3	1	1/3	Simple Multiple Choice – Computer Scored
R456Q02	DR456Q02C	PR456Q02	Biscuits	PB and CB	R04, U3	2	2/3	Open Response – Human Coded
R456Q06	DR456Q06C	PR456Q06	Biscuits	PB and CB	R04, U3	3	3/3	Open Response – Human Coded
R460Q01	DR460Q01C	R460Q01	Gulf of Mexico	PB and CB	R04	12	1/3	Open Response – Human Coded
R460Q05	CR460Q05S	PR460Q05S	Gulf of Mexico	PB and CB	R04	13	2/3	Simple Multiple Choice – Computer Scored
R460Q06	CR460Q06S	PR460Q06S	Gulf of Mexico	PB and CB	R04	14	3/3	Simple Multiple Choice – Computer Scored
R462Q02	DR462Q02C	R462Q02	Parcel Post	PB and CB	R06B	5	1/3	Open Response – Human Coded
R462Q04	CR462Q04S	PR462Q04S	Parcel Post	PB and CB	R06B	7	3/3	Simple Multiple Choice – Computer Scored
R462Q05	DR462Q05C	R462Q05	Parcel Post	PB and CB	R06B	6	2/3	Open Response – Human Coded
R465Q01	DR465Q01C	R465Q01	How to survive at work	PB and CB	R06B	12	1/4	Open Response – Human Coded
R465Q02	DR465Q02C	R465Q02	How to survive at work	PB and CB	R06B	13	2/4	Open Response – Human Coded
R465Q05	DR465Q05C	R465Q05	How to survive at work	PB and CB	R06B	15	4/4	Open Response – Human Coded
R465Q06	DR465Q06C	R465Q06	How to survive at work	PB and CB	R06B	14	3/4	Open Response – Human Coded
R466Q02	DR466Q02C	PR466Q02	Work Right	PB and CB	R04	4	1/3	Open Response – Human Coded
R466Q03	CR466Q03S	PR466Q03S	Work Right	PB and CB	R04	5	2/3	Complex Multiple Choice – Computer Scored
R466Q06	CR466Q06S	PR466Q06S	Work Right	PB and CB	R04	6	3/3	Open Response – Computer Scored

## [Part 3/8]

Table A.3 PISA 2015 main survey trend reading item classification

Generic ID	Item format - PBA	Situation	Text format	Aspect	Unit origin	Language of submission	Source
R055Q01	Simple Multiple Choice – Data Entered	Public	Continuous	Integrate and interpret	CITO	English	2009
R055Q02	Open Response – Human Coded	Public	Continuous	Reflect and evaluate	CITO	English	2009
R055Q03	Open Response – Human Coded	Public	Continuous	Integrate and interpret	CITO	English	2009
R055Q05	Open Response – Human Coded	Public	Continuous	Integrate and interpret	CITO	English	2009
R067Q01	Simple Multiple Choice – Data Entered	Personal	Continuous	Integrate and interpret	Greece	Greek	2009
R067Q04	Open Response – Human Coded	Personal	Continuous	Reflect and evaluate	Greece	Greek	2009
R067Q05	Open Response – Human Coded	Personal	Continuous	Reflect and evaluate	Greece	Greek	2009
R083Q01	Simple Multiple Choice – Data Entered	Educational	Mixed	Integrate and interpret	ACER	English	2009
R083Q02	Open Response – Human Coded	Educational	Non-continuous	Access and retrieve	ACER	English	2009
R083Q03	Open Response – Human Coded	Educational	Non-continuous	Access and retrieve	ACER	English	2009
R083Q04	Simple Multiple Choice – Data Entered	Educational	Non-continuous	Integrate and interpret	ACER	English	2009
R101Q01	Simple Multiple Choice – Data Entered	Public	Continuous	Integrate and interpret	Sweden	Swedish	2009
R101Q02	Simple Multiple Choice – Data Entered	Public	Continuous	Integrate and interpret	Sweden	Swedish	2009
R101Q03	Simple Multiple Choice – Data Entered	Public	Continuous	Reflect and evaluate	Sweden	Swedish	2009
R101Q04	Simple Multiple Choice – Data Entered	Public	Continuous	Integrate and interpret	Sweden	Swedish	2009
R101Q05	Simple Multiple Choice – Data Entered	Public	Continuous	Integrate and interpret	Sweden	Swedish	2009
R102Q04	Open Response – Human Coded	Personal	Continuous	Integrate and interpret	CITO	English	2009
R102Q05	Open Response – Human Coded	Personal	Non-continuous	Integrate and interpret	CITO	English	2009
R102Q07	Simple Multiple Choice – Data Entered	Personal	Mixed	Integrate and interpret	CITO	English	2009
R104Q01	Open Response – Human Coded	Public	Non-continuous	Access and retrieve	New Zealand	English	2009
R104Q02	Open Response – Human Coded	Public	Non-continuous	Access and retrieve	New Zealand	English	2009
R104Q05	Open Response – Human Coded	Public	Non-continuous	Access and retrieve	New Zealand	English	2009
R111Q01	Simple Multiple Choice – Data Entered	Educational	Continuous	Integrate and interpret	Finland	Finnish	2009
R111Q02	Open Response – Human Coded	Educational	Continuous	Reflect and evaluate	Finland	Finnish	2009
R111Q06	Open Response – Human Coded	Educational	Continuous	Reflect and evaluate	Finland	Finnish	2009
R219Q01	Open Response – Human Coded	Occupational	Non-continuous	Access and retrieve	IALS	IALS	2009
R219Q01	Open Response – Human Coded	Occupational	Non-continuous	Integrate and interpret	IALS	IALS	2009
R219Q02	Open Response – Human Coded	Occupational	Non-continuous	Reflect and evaluate	IALS	IALS	2009
R220Q01	Open Response – Human Coded	Educational	Mixed	Access and retrieve	France	French	2009-2012
R220Q02	Simple Multiple Choice – Data Entered	Educational	Mixed	Integrate and interpret	France	French	2009-2012
R220Q04	Simple Multiple Choice – Data Entered	Educational	Continuous	Integrate and interpret	France	French	2009-2012
R220Q05	Simple Multiple Choice – Data Entered	Educational	Continuous	Integrate and interpret	France	French	2009
R220Q06	Simple Multiple Choice – Data Entered	Educational	Continuous	Integrate and interpret	France	French	2009
R227Q01	Simple Multiple Choice – Data Entered	Occupational	Mixed	Integrate and interpret	Switzerland	German	2009
R227Q02	Complex Multiple Choice – Data Entered	Occupational	Continuous	Access and retrieve	Switzerland	German	2009
R227Q03	Open Response – Human Coded	Occupational	Continuous	Reflect and evaluate	Switzerland	German	2009
R227Q06	Open Response – Human Coded	Occupational	Non-continuous	Access and retrieve	Switzerland	German	2009
R245Q01	Open Response – Human Coded	Personal	Multiple	Access and retrieve	IALS	English	2009
R245Q02	Open Response – Human Coded	Personal	Multiple	Integrate and interpret	IALS	English	2009
R404Q03	Simple Multiple Choice – Data Entered	Public	Continuous	Integrate and interpret	ILS	Norwegian	2012
R404Q06	Simple Multiple Choice – Data Entered	Public	Non-continuous	Integrate and interpret	ILS	Norwegian	2012
R404Q07	Complex Multiple Choice – Data Entered	Public	Non-continuous	Integrate and interpret	ILS	Norwegian	2012
R404Q10	Open Response – Human Coded	Public	Non-continuous	Reflect and evaluate	ILS	Norwegian	2012
R404Q10	Open Response – Human Coded	Public	Non-continuous	Reflect and evaluate	ILS	Norwegian	2012
R406Q01	Open Response – Human Coded	Personal	Continuous	Integrate and interpret	NIER	Japanese	2012
R406Q02	Open Response – Human Coded	Personal	Continuous	Integrate and interpret	NIER	Japanese	2012
R406Q05	Open Response – Human Coded	Personal	Continuous	Integrate and interpret	NIER	Japanese	2012
R412Q01	Simple Multiple Choice – Data Entered	Educational	Non-continuous	Access and retrieve	ACER	English	2012
R412Q05	Simple Multiple Choice – Data Entered	Educational	Continuous	Integrate and interpret	ACER	English	2012
R412Q06	Complex Multiple Choice – Data Entered	Educational	Continuous	Integrate and interpret	ACER	English	2012
R412Q08	Open Response – Human Coded	Educational	Mixed	Integrate and interpret	ACER	English	2012
R420Q02	Open Response – Human Coded	Educational	Non-continuous	Access and retrieve	NIER	Japanese	2012



## [Part 4/8]

Table A.3 PISA 2015 main survey trend reading item classification

Generic ID	Item format - PBA	Situation	Text format	Aspect	Unit origin	Language of submission	Source
R420Q06	Open Response – Human Coded	Educational	Non-continuous	Reflect and evaluate	NIER	Japanese	2012
R420Q09	Open Response – Human Coded	Educational	Non-continuous	Access and retrieve	NIER	Japanese	2012
R420Q10	Open Response – Human Coded	Educational	Non-continuous	Integrate and interpret	NIER	Japanese	2012
R424Q02	Complex Multiple Choice – Data Entered	Educational	Non-continuous	Integrate and interpret	aSPe	French	2012
R424Q03	Simple Multiple Choice – Data Entered	Educational	Non-continuous	Reflect and evaluate	aSPe	French	2012
R424Q07	Simple Multiple Choice – Data Entered	Educational	Continuous	Reflect and evaluate	aSPe	French	2012
R432Q01	Open Response – Human Coded	Personal	Continuous	Integrate and interpret	DIPF	German	2012
R432Q05	Open Response – Human Coded	Personal	Multiple	Reflect and evaluate	DIPF	German	2012
R432Q06	Complex Multiple Choice – Data Entered	Personal	Continuous	Integrate and interpret	DIPF	German	2012
R435Q01	Simple Multiple Choice – Data Entered	Educational	Continuous	Integrate and interpret	Canada	English	2009
R435Q02	Open Response – Human Coded	Educational	Continuous	Access and retrieve	Canada	English	2009
R435Q05	Open Response – Human Coded	Educational	Continuous	Reflect and evaluate	Canada	English	2009
R435Q08	Complex Multiple Choice – Data Entered	Educational	Continuous	Reflect and evaluate	Canada	English	2009
R437Q01	Simple Multiple Choice – Data Entered	Personal	Continuous	Integrate and interpret	Sweden	Portuguese	2012
R437Q06	Simple Multiple Choice – Data Entered	Personal	Continuous	Integrate and interpret	Sweden	Portuguese	2012
R437Q07	Open Response – Human Coded	Personal	Continuous	Integrate and interpret	Sweden	Portuguese	2012
R442Q02	Open Response – Human Coded	Personal	Continuous	Access and retrieve	Colombia	Spanish	2009
R442Q03	Open Response – Human Coded	Personal	Continuous	Integrate and interpret	Colombia	Spanish	2009
R442Q05	Open Response – Human Coded	Personal	Continuous	Reflect and evaluate	Colombia	Spanish	2009
R442Q06	Open Response – Human Coded	Personal	Continuous	Reflect and evaluate	Colombia	Spanish	2009
R442Q07	Simple Multiple Choice – Data Entered	Personal	Continuous	Integrate and interpret	Colombia	Spanish	2009
R445Q01	Open Response – Human Coded	Public	Continuous	Integrate and interpret	Spain	Spanish	2009
R445Q03	Simple Multiple Choice – Data Entered	Public	Continuous	Integrate and interpret	Spain	Spanish	2009
R445Q04	Simple Multiple Choice – Data Entered	Public	Continuous	Integrate and interpret	Spain	Spanish	2009
R445Q06	Simple Multiple Choice – Data Entered	Public	Continuous	Integrate and interpret	Spain	Spanish	2009
R446Q03	Open Response – Human Coded	Occupational	Non-continuous	Access and retrieve	ACER	English	2012
R446Q06	Open Response – Human Coded	Occupational	Non-continuous	Reflect and evaluate	ACER	English	2012
R453Q01	Simple Multiple Choice – Data Entered	Occupational	Continuous	Integrate and interpret	Finland	Finnish	2012
R453Q04	Open Response – Human Coded	Occupational	Continuous	Reflect and evaluate	Finland	Finnish	2012
R453Q05	Complex Multiple Choice – Data Entered	Occupational	Continuous	Access and retrieve	Finland	Finnish	2012
R453Q06	Open Response – Human Coded	Occupational	Continuous	Reflect and evaluate	Finland	Finnish	2012
R455Q02	Open Response – Human Coded	Personal	Continuous	Reflect and evaluate	New Zealand	English	2012
R455Q03	Open Response – Human Coded	Personal	Continuous	Access and retrieve	New Zealand	English	2012
R455Q04	Simple Multiple Choice – Data Entered	Personal	Continuous	Integrate and interpret	New Zealand	English	2012
R455Q05	Complex Multiple Choice – Data Entered	Personal	Continuous	Integrate and interpret	New Zealand	English	2012
R456Q01	Simple Multiple Choice – Data Entered	Personal	Continuous	Access and retrieve	Serbia	English	2012
R456Q02	Open Response – Human Coded	Personal	Continuous	Integrate and interpret	Serbia	English	2012
R456Q06	Open Response – Human Coded	Personal	Continuous	Integrate and interpret	Serbia	English	2012
R460Q01	Open Response – Human Coded	Educational	Continuous	Access and retrieve	Mexico	Spanish	2009
R460Q05	Simple Multiple Choice – Data Entered	Educational	Continuous	Access and retrieve	Mexico	Spanish	2009
R460Q06	Simple Multiple Choice – Data Entered	Educational	Continuous	Integrate and interpret	Mexico	Spanish	2009
R462Q02	Open Response – Human Coded	Public	Non-continuous	Access and retrieve	Greece	Greek	2009
R462Q04	Simple Multiple Choice – Data Entered	Public	Non-continuous	Access and retrieve	Greece	Greek	2009
R462Q05	Open Response – Human Coded	Public	Non-continuous	Integrate and interpret	Greece	Greek	2009
R465Q01	Open Response – Human Coded	Occupational	Non-continuous	Access and retrieve	ACER	English	2009
R465Q02	Open Response – Human Coded	Occupational	Non-continuous	Integrate and interpret	ACER	English	2009
R465Q05	Open Response – Human Coded	Occupational	Non-continuous	Reflect and evaluate	ACER	English	2009
R465Q06	Open Response – Human Coded	Occupational	Non-continuous	Reflect and evaluate	ACER	English	2009
R466Q02	Open Response – Human Coded	Occupational	Continuous	Access and retrieve	aSPe	French	2012
R466Q03	Complex Multiple Choice – Data Entered	Occupational	Mixed	Integrate and interpret	aSPe	French	2012
R466Q06	Open Response – Human Coded	Occupational	Continuous	Access and retrieve	aSPe	French	2012

## [Part 5/8]

Table A.3 PISA 2015 main survey trend reading item classification

Generic ID	CBA international % correct	CBA international % correct S.E.	CBA item parameters (RP = 0.50)				CBA thresholds (RP = 0.62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
R055Q01	76.02	0.22	1.000	-0.629			450		Level 2
R055Q02	44.05	0.25	1.011	0.462			593		Level 4
R055Q03	50.99	0.23	1.237	0.453			585		Level 4
R055Q05	65.23	0.25	1.417	-0.135			504		Level 3
R067Q01	84.79	0.18	1.000	-1.223			372		Level 1a
R067Q04	55.11	0.22	0.540	0.054	0.154	-0.154	502	635	Level 5
R067Q05	66.00	0.23	0.592	-0.187	-0.964	0.964	470	533	Level 3
R083Q01	59.15	0.30	1.150	0.166			549		Level 3
R083Q02	77.80	0.25	1.000	-0.545			461		Level 2
R083Q03	73.33	0.26	1.000	-0.341			488		Level 3
R083Q04	65.37	0.28	0.713	-0.179			524		Level 3
R101Q01	52.18	0.29	0.823	0.452			600		Level 4
R101Q02	83.96	0.22	1.000	-0.799			427		Level 2
R101Q03	64.16	0.29	1.131	0.084			539		Level 3
R101Q04	77.16	0.25	1.094	-0.528			460		Level 2
R101Q05	43.47	0.29	0.632	0.767			655		Level 5
R102Q04	22.13	0.21	1.149	1.214			687		Level 5
R102Q05	31.63	0.24	1.052	0.862			644		Level 5
R102Q07	80.43	0.21	1.138	-0.785			424		Level 2
R104Q01	53.74	0.26	1.182	0.219			555		Level 4
R104Q02	38.24	0.26	0.535	1.141			716		Level 6
R104Q05	11.26	0.12	0.941	2.068	0.628	-0.628	767	896	Level 6
R111Q01	63.43	0.25	1.024	-0.112			517		Level 3
R111Q02	33.42	0.19	0.791	0.876	0.427	-0.427	610	729	Level 6
R111Q06	37.83	0.22	0.724	0.663	-0.564	0.564	582	642	Level 5
R219Q01	66.06	0.23	1.518	-0.002			519		Level 3
R219Q01	59.30	0.24	1.341	-0.009			522		Level 3
R219Q02	71.89	0.23	1.175	-0.488			463		Level 2
R220Q01	18.17	0.20	1.237	1.273			693		Level 5
R220Q02	49.46	0.26	1.000	0.311			573		Level 4
R220Q04	58.17	0.26	0.857	0.042			544		Level 3
R220Q05	78.55	0.21	1.000	-0.741			435		Level 2
R220Q06	59.42	0.26	0.774	-0.076			534		Level 3
R227Q01	54.82	0.25	0.599	0.080			568		Level 4
R227Q02	44.69	0.18	0.779	0.539	0.677	-0.677	565	712	Level 6
R227Q03	52.12	0.26	1.151	0.245			560		Level 4
R227Q06	66.08	0.25	1.355	-0.175			499		Level 3
R245Q01	60.82	0.29	0.979	0.004			534		Level 3
R245Q02	64.03	0.28	1.085	-0.006			529		Level 3
R404Q03	70.55	0.23	1.000	-0.455			473		Level 2
R404Q06	43.56	0.26	0.701	0.491			613		Level 4
R404Q07	28.70	0.23	0.899	1.156			689		Level 5
R404Q10	43.69	0.26	1.561	0.548			591		Level 4
R404Q10	36.72	0.25	1.454	0.720			615		Level 4
R406Q01	60.25	0.26	1.033	-0.022			528		Level 3
R406Q02	31.33	0.23	0.762	1.086			687		Level 5
R406Q05	66.75	0.24	1.000	-0.321			490		Level 3
R412Q01	81.54	0.20	1.000	-0.955			407		Level 1a
R412Q05	53.84	0.26	0.719	0.119			563		Level 4
R412Q06	38.22	0.27	0.147	2.656			1102		Level 6
R412Q08	34.96	0.25	1.255	0.701			617		Level 4
R420Q02	73.12	0.23	1.137	-0.489			464		Level 2



[Part 6/8]

Table A.3 PISA 2015 main survey trend reading item classification

Generic ID	CBA international % correct	CBA international % correct S.E.	CBA item parameters (RP = 0.50)				CBA thresholds (RP = 0.62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
R420Q06	40.52	0.25	0.725	0.660			634		Level 5
R420Q09	71.88	0.24	1.000	-0.428			476		Level 2
R420Q10	59.83	0.24	1.000	0.076	-1.373	1.373	505	528	Level 3
R424Q02	36.36	0.25	0.533	0.837			676		Level 5
R424Q03	50.57	0.25	0.572	0.266			596		Level 4
R424Q07	73.80	0.23	1.000	-0.525			463		Level 2
R432Q01	80.37	0.20	1.481	-0.648			435		Level 2
R432Q05	65.79	0.24	1.399	-0.122			506		Level 3
R432Q06	8.12	0.14	1.000	2.104			809		Level 6
R435Q01	65.55	0.50	0.825	-0.646			456		Level 2
R435Q02	82.93	0.38	1.000	-1.430			344		Level 1a
R435Q05	60.39	0.52	1.000	-0.440			474		Level 2
R435Q08	51.90	0.50	1.169	-0.112			512		Level 3
R437Q01	43.90	0.27	0.563	0.585			639		Level 5
R437Q06	53.30	0.26	0.685	0.220			579		Level 4
R437Q07	16.90	0.20	0.774	1.775			777		Level 6
R442Q02	72.09	0.27	0.973	-0.297			494		Level 3
R442Q03	70.37	0.27	1.791	-0.057			508		Level 3
R442Q05	32.46	0.28	1.669	0.843			628		Level 5
R442Q06	22.41	0.25	1.546	1.227			681		Level 5
R442Q07	39.11	0.29	1.454	0.757			620		Level 4
R445Q01	68.35	0.47	0.738	-0.565			471		Level 2
R445Q03	79.85	0.43	1.074	-1.279			362		Level 1a
R445Q04	82.35	0.38	0.955	-1.291			364		Level 1a
R445Q06	58.19	0.51	0.843	-0.459			479		Level 2
R446Q03	88.45	0.16	1.000	-1.292			362		Level 1a
R446Q06	68.14	0.24	1.123	-0.250			495		Level 3
R453Q01	77.33	0.22	1.172	-0.553			454		Level 2
R453Q04	61.20	0.25	0.994	-0.059			525		Level 3
R453Q05	51.67	0.25	1.172	0.229			557		Level 4
R453Q06	66.64	0.24	1.438	-0.202			494		Level 3
R455Q02	31.95	0.24	0.762	1.039			681		Level 5
R455Q03	76.57	0.23	0.917	-0.665			448		Level 2
R455Q04	57.98	0.26	0.844	0.028			543		Level 3
R455Q05	21.65	0.21	1.140	1.277			696		Level 5
R456Q01	94.41	0.12	1.000	-1.912			281		Level 1b
R456Q02	72.62	0.23	1.000	-0.496			467		Level 2
R456Q06	76.48	0.23	1.000	-0.655			446		Level 2
R460Q01	63.88	0.25	1.145	-0.115			513		Level 3
R460Q05	77.09	0.22	1.000	-0.635			449		Level 2
R460Q06	60.08	0.26	0.826	0.027			544		Level 3
R462Q02	37.35	0.49	1.016	0.310			573		Level 4
R462Q04	64.69	0.48	0.721	-0.884			431		Level 2
R462Q05	36.72	0.50	1.078	0.210			557		Level 4
R465Q01	87.60	0.34	1.139	-1.616			315		Level 1b
R465Q02	48.36	0.54	1.539	0.050			526		Level 3
R465Q05	43.77	0.54	1.086	0.203			556		Level 4
R465Q06	57.84	0.54	1.055	-0.157			510		Level 3
R466Q02	41.40	0.25	1.293	0.536			594		Level 4
R466Q03	13.00	0.17	1.000	1.621			746		Level 6
R466Q06	78.34	0.21	1.466	-0.429			464		Level 2

## [Part 7/8]

Table A.3 PISA 2015 main survey trend reading item classification

Generic ID	PBA international % correct	PBA international % correct S.E.	PBA item parameters (RP = 0.50)				PBA thresholds (RP = 0.62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
R055Q01	61.37	0.54	1.000	-0.752			433		Level 2
R055Q02	35.11	0.52	1.011	0.462			593		Level 4
R055Q03	36.65	0.52	1.237	0.161			546		Level 3
R055Q05	48.82	0.58	1.417	-0.280			484		Level 3
R067Q01	75.85	0.47	1.000	-1.223			372		Level 1a
R067Q04	37.55	0.42	0.540	0.054	0.154	-0.154	502	635	Level 5
R067Q05	39.43	0.43	0.592	-0.187	-0.964	0.964	470	533	Level 3
R083Q01	40.52	0.89	1.150	0.166			549		Level 3
R083Q02	69.01	1.08	1.000	-0.697			441		Level 2
R083Q03	62.64	0.98	1.000	-0.575			457		Level 2
R083Q04	62.17	1.04	0.713	-0.179			524		Level 3
R101Q01	30.01	1.02	0.823	0.452			600		Level 4
R101Q02	75.06	0.99	1.000	-0.799			427		Level 2
R101Q03	44.75	1.12	1.131	0.084			539		Level 3
R101Q04	69.17	1.04	1.094	-0.528			460		Level 2
R101Q05	34.04	1.04	0.632	0.644			639		Level 5
R102Q04	11.57	0.31	1.149	1.074			669		Level 5
R102Q05	20.31	0.42	1.052	0.710			624		Level 4
R102Q07	62.59	0.54	1.138	-0.785			424		Level 2
R104Q01	56.01	0.60	1.182	-0.565			452		Level 2
R104Q02	23.33	0.48	0.535	1.141			716		Level 6
R104Q05	12.42	0.29	0.941	1.516	0.709	-0.709	694	832	Level 6
R111Q01	40.87	0.55	1.024	-0.112			517		Level 3
R111Q02	23.20	0.39	0.791	0.876	0.427	-0.427	610	729	Level 6
R111Q06	20.19	0.40	0.724	0.663	-0.564	0.564	582	642	Level 5
R219Q01	30.47	0.47	1.518	-0.002			519		Level 3
R219Q01	25.19	0.44	1.341	0.418			578		Level 4
R219Q02	51.10	0.50	1.175	-0.488			463		Level 2
R220Q01	16.15	0.39	1.237	0.752			624		Level 4
R220Q02	43.14	0.55	1.000	0.050			539		Level 3
R220Q04	42.32	0.59	0.857	0.042			544		Level 3
R220Q05	60.27	0.57	1.000	-0.741			435		Level 2
R220Q06	49.67	0.58	0.774	-0.262			509		Level 3
R227Q01	33.39	0.48	0.599	0.434			615		Level 4
R227Q02	30.91	0.36	0.779	0.379	0.768	-0.768	544	701	Level 6
R227Q03	36.02	0.50	1.151	0.245			560		Level 4
R227Q06	45.69	0.54	1.355	-0.175			499		Level 3
R245Q01	53.37	1.06	0.979	-0.122			517		Level 3
R245Q02	55.31	1.03	1.085	-0.151			510		Level 3
R404Q03	59.66	0.56	1.000	-0.455			473		Level 2
R404Q06	34.63	0.51	0.701	0.491			613		Level 4
R404Q07	18.79	0.41	0.899	1.093			681		Level 5
R404Q10	23.43	0.45	1.561	0.548			591		Level 4
R404Q10	23.70	0.45	1.454	0.720			615		Level 4
R406Q01	50.32	0.53	1.033	-0.198			505		Level 3
R406Q02	27.23	0.46	0.762	0.946			669		Level 5
R406Q05	55.98	0.51	1.000	-0.418			477		Level 2
R412Q01	74.80	0.45	1.000	-0.955			407		Level 1a
R412Q05	41.99	0.55	0.719	0.119			563		Level 4
R412Q06	35.77	0.67	0.147	2.656			1102		Level 6
R412Q08	21.86	0.42	1.255	0.701			617		Level 4
R420Q02	58.47	0.53	1.137	-0.658			441		Level 2



[Part 8/8]

Table A.3 PISA 2015 main survey trend reading item classification

Generic ID	PBA international % correct	PBA international % correct S.E.	PBA item parameters (RP = 0.50)				PBA thresholds (RP = 0.62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
R420Q06	28.39	0.46	0.725	0.660			634		Level 5
R420Q09	61.86	0.51	1.000	-0.590			455		Level 2
R420Q10	44.62	0.53	1.000	-0.098	-1.125	1.125	482	508	Level 3
R424Q02	31.48	0.48	0.533	0.837			676		Level 5
R424Q03	50.89	0.52	0.572	-0.284			523		Level 3
R424Q07	58.49	0.56	1.000	-0.525			463		Level 2
R432Q01	67.08	0.49	1.481	-0.648			435		Level 2
R432Q05	44.68	0.51	1.399	-0.122			506		Level 3
R432Q06	6.70	0.27	1.000	1.750			763		Level 6
R435Q01	60.54	0.55	0.825	-0.646			456		Level 2
R435Q02	84.24	0.45	1.000	-1.574			325		Level 1b
R435Q05	58.50	0.61	1.000	-0.440			474		Level 2
R435Q08	45.27	0.62	1.169	-0.112			512		Level 3
R437Q01	41.27	0.53	0.563	0.342			607		Level 4
R437Q06	40.96	0.52	0.685	0.220			579		Level 4
R437Q07	19.05	0.42	0.774	1.775			777		Level 6
R442Q02	67.61	0.99	0.973	-0.297			494		Level 3
R442Q03	52.31	1.10	1.791	-0.057			508		Level 3
R442Q05	21.46	0.85	1.669	0.843			628		Level 5
R442Q06	21.04	0.88	1.546	1.073			660		Level 5
R442Q07	20.49	0.82	1.454	0.757			620		Level 4
R445Q01	57.56	0.58	0.738	-0.565			471		Level 2
R445Q03	75.65	0.53	1.074	-1.279			362		Level 1a
R445Q04	75.64	0.50	0.955	-1.291			364		Level 1a
R445Q06	54.62	0.58	0.843	-0.459			479		Level 2
R446Q03	77.51	0.45	1.000	-1.292			362		Level 1a
R446Q06	48.46	0.51	1.123	-0.412			474		Level 2
R453Q01	57.45	0.53	1.172	-0.553			454		Level 2
R453Q04	44.99	0.55	0.994	0.008			534		Level 3
R453Q05	34.49	0.50	1.172	0.081			538		Level 3
R453Q06	40.60	0.52	1.438	-0.118			505		Level 3
R455Q02	19.91	0.41	0.762	1.039			681		Level 5
R455Q03	66.79	0.52	0.917	-0.665			448		Level 2
R455Q04	42.15	0.53	0.844	-0.084			528		Level 3
R455Q05	11.50	0.31	1.140	1.276			696		Level 5
R456Q01	90.38	0.31	1.000	-1.912			281		Level 1b
R456Q02	67.36	0.48	1.000	-0.752			434		Level 2
R456Q06	63.77	0.50	1.000	-0.790			428		Level 2
R460Q01	55.21	0.55	1.145	-0.115			513		Level 3
R460Q05	66.84	0.51	1.000	-0.757			433		Level 2
R460Q06	45.67	0.57	0.826	0.027			544		Level 3
R462Q02	37.19	0.59	1.016	0.310			573		Level 4
R462Q04	64.05	0.57	0.721	-0.884			431		Level 2
R462Q05	38.88	0.59	1.078	0.210			557		Level 4
R465Q01	80.73	0.48	1.139	-1.616			315		Level 1b
R465Q02	35.90	0.63	1.539	0.050			526		Level 3
R465Q05	45.50	0.62	1.086	0.203			556		Level 4
R465Q06	41.99	0.58	1.055	-0.157			510		Level 3
R466Q02	31.50	0.51	1.293	0.536			594		Level 4
R466Q03	11.71	0.31	1.000	1.621			746		Level 6
R466Q06	56.00	0.54	1.466	-0.429			464		Level 2

## [Part 1/8]

Table A.4 PISA 2015 main survey trend math item classification

Generic ID	CBA item ID in main survey analysis output	PBA item ID in main survey analysis output	Unit name	Mode [paper-based (PB); computer-based (CB)]	2015 field trial and main survey cluster	Sequence in cluster	Sequence in unit	Item format – CBA
M033Q01	CM033Q01S	PM033Q01S	A View Room	PB and CB	M1	1	1/1	Simple Multiple Choice – Computer Scored
M474Q01	CM474Q01S	PM474Q01	Running Time	PB and CB	M1	2	1/1	Simple Multiple Choice – Computer Scored
M155Q02	DM155Q02C	PM155Q02	Population Pyramids	PB and CB	M1	3	1/4	Open Response – Human Coded
M155Q01	CM155Q01S	PM155Q01	Population Pyramids	PB and CB	M1	4	2/4	Complex Multiple Choice – Computer Scored
M155Q03	DM155Q03C	PM155Q03	Population Pyramids	PB and CB	M1	5	3/4	Open Response – Human Coded
M155Q04	CM155Q04S	PM155Q04S	Population Pyramids	PB and CB	M1	6	4/4	Complex Multiple Choice – Computer Scored
M411Q01	CM411Q01S	PM411Q01	Diving	PB and CB	M1	7	1/2	Open Response – Computer Scored
M411Q02	CM411Q02S	PM411Q02S	Diving	PB and CB	M1	8	2/2	Simple Multiple Choice – Computer Scored
M803Q01	CM803Q01S	PM803Q01S	Labels	PB and CB	M1	9	1/1	Open Response – Computer Scored
M442Q02	CM442Q02S	PM442Q02	Braille	PB and CB	M1	10	1/1	Complex Multiple Choice – Computer Scored
M462Q01	DM462Q01C	PM462Q01	Third Side	PB and CB	M1	11	1/1	Open Response – Human Coded
M034Q01	CM034Q01S	PM034Q01S	Bricks	PB and CB	M1	12	1/1	Open Response – Computer Scored
M305Q01	CM305Q01S	PM305Q01S	Map	PB and CB	M2	1	1/1	Simple Multiple Choice – Computer Scored
M496Q01	CM496Q01S	PM496Q01S	Cash Withdrawal	PB and CB	M2	2	1/2	Complex Multiple Choice – Computer Scored
M496Q02	CM496Q02S	PM496Q02	Cash Withdrawal	PB and CB	M2	3	2/2	Open Response – Computer Scored
M423Q01	CM423Q01S	PM423Q01S	Tossing Coins	PB and CB	M2	4	1/1	Simple Multiple Choice – Computer Scored
M406Q01	DM406Q01C	PM406Q01	Running Tracks	PB and CB	M2	6	1/2	Open Response – Human Coded
M406Q02	DM406Q02C	PM406Q02	Running Tracks	PB and CB	M2	7	2/2	Open Response – Human Coded
M603Q01	CM603Q01S	PM603Q01S	Number Check	PB and CB	M2	8	1/1	Complex Multiple Choice – Computer Scored
M571Q01	CM571Q01S	PM571Q01S	Stop The Car	PB and CB	M2	9	1/1	Simple Multiple Choice – Computer Scored
M564Q01	CM564Q01S	PM564Q01S	Chair Lift	PB and CB	M2	10	1/2	Simple Multiple Choice – Computer Scored
M564Q02	CM564Q02S	PM564Q02S	Chair Lift	PB and CB	M2	11	2/2	Simple Multiple Choice – Computer Scored
M447Q01	CM447Q01S	PM447Q01S	Tile Arrangement	PB and CB	M3	1	1/1	Simple Multiple Choice – Computer Scored
M273Q01	CM273Q01S	PM273Q01S	Pipelines	PB and CB	M3	2	1/1	Complex Multiple Choice – Computer Scored
R111Q06	CM408Q01S	PM408Q01S	Lotteries	PB and CB	M3	3	1/1	Complex Multiple Choice – Computer Scored
M420Q01	CM420Q01S	PM420Q01S	Transport	PB and CB	M3	4	1/1	Complex Multiple Choice – Computer Scored
M446Q01	CM446Q01S	PM446Q01	Thermometer Cricket	PB and CB	M3	5	1/2	Open Response – Computer Scored
M446Q02	DM446Q02C	PM446Q02	Thermometer Cricket	PB and CB	M3	6	2/2	Open Response – Human Coded
M559Q01	CM559Q01S	PM559Q01S	Telephone Rates	PB and CB	M3	7	1/1	Simple Multiple Choice – Computer Scored
M828Q02	DM828Q02C	PM828Q02	Carbon Dioxide	PB and CB	M3	9	1/2	Open Response – Human Coded
M828Q03	CM828Q03S	PM828Q03	Carbon Dioxide	PB and CB	M3	10	2/2	Open Response – Computer Scored
M464Q01	CM464Q01S	PM464Q01S	Fence	PB and CB	M3	11	1/1	Open Response – Computer Scored
M800Q01	CM800Q01S	PM800Q01S	Computer Game	PB and CB	M3	12	1/1	Simple Multiple Choice – Computer Scored
M982Q01	CM982Q01S	PM982Q01	Employment Data	PB and CB	M4	1	1/4	Open Response – Computer Scored
M982Q02	CM982Q02S	PM982Q02	Employment Data	PB and CB	M4	2	2/4	Open Response – Computer Scored
M982Q03	CM982Q03S	PM982Q03S	Employment Data	PB and CB	M4	3	3/4	Complex Multiple Choice – Computer Scored
M982Q04	CM982Q04S	PM982Q04S	Employment Data	PB and CB	M4	4	4/4	Simple Multiple Choice – Computer Scored
M992Q01	CM992Q01S	PM992Q01	Spacers	PB and CB	M4	5	1/3	Open Response – Computer Scored
M992Q02	CM992Q02S	PM992Q02	Spacers	PB and CB	M4	6	2/3	Open Response – Computer Scored
M992Q03	DM992Q03C	PM992Q03	Spacers	PB and CB	M4	7	3/3	Open Response – Human Coded
M915Q01	CM915Q01S	PM915Q01S	Carbon Tax	PB and CB	M4, U4	8	1/2	Simple Multiple Choice – Computer Scored
M915Q02	CM915Q02S	PM915Q02	Carbon Tax	PB and CB	M4, U4	9	2/2	Open Response – Computer Scored



[Part 2/8]

Table A.4 PISA 2015 main survey trend math item classification

Generic ID	CBA item ID in main survey analysis output	PBA item ID in main survey analysis output	Unit name	Mode [paper-based (PB); computer-based (CB)]	2015 field trial and main survey cluster	Sequence in cluster	Sequence in unit	Item format – CBA
M906Q01	CM906Q01S	PM906Q01S	Crazy Ants	PB and CB	M4	10	1/2	Simple Multiple Choice – Computer Scored
M906Q02	DM906Q02C	PM906Q02	Crazy Ants	PB and CB	M4	11	2/2	Open Response – Human Coded
M00KQ02	DM00KQ02C	PM00KQ02	Wheelchair Basketball	PB and CB	M4	12	1/1	Open Response – Human Coded
M909Q01	CM909Q01S	PM909Q01	Speeding Fines	PB and CB	M5, U4	1	1/3	Open Response – Computer Scored
M909Q02	CM909Q02S	PM909Q02S	Speeding Fines	PB and CB	M5, U4	2	2/3	Simple Multiple Choice – Computer Scored
M909Q03	CM909Q03S	PM909Q03	Speeding Fines	PB and CB	M5, U4	3	3/3	Open Response – Computer Scored
M949Q01	CM949Q01S	PM949Q01S	Roof Truss Design	PB and CB	M5, U4	4	1/3	Complex Multiple Choice – Computer Scored
M949Q02	CM949Q02S	PM949Q02S	Roof Truss Design	PB and CB	M5, U4	5	2/3	Complex Multiple Choice – Computer Scored
M949Q03	DM949Q03C	PM949Q03	Roof Truss Design	PB and CB	M5, U4	6	3/3	Open Response – Human Coded
M00GQ01	CM00GQ01S	PM00GQ01	Advertising Column	PB and CB	M5	7	1/1	Open Response – Computer Scored
M955Q01	DM955Q01C	PM955Q01	Migration	PB and CB	M5	8	1/3	Open Response – Human Coded
M955Q02	DM955Q02C	PM955Q02	Migration	PB and CB	M5	9	2/3	Open Response – Human Coded
M955Q03	CM955Q03S	PM955Q03	Migration	PB and CB	M5	10	3/3	Open Response – Computer Scored
M998Q02	DM998Q02C	PM998Q02	Bike Rental	PB and CB	M5	11	1/2	Open Response – Human Coded
M998Q04	CM998Q04S	PM998Q04S	Bike Rental	PB and CB	M5	12	2/2	Complex Multiple Choice – Computer Scored
M905Q01	CM905Q01S	PM905Q01S	Tennis balls	PB and CB	M6A	1	1/2	Complex Multiple Choice – Computer Scored
M905Q02	DM905Q02C	PM905Q02	Tennis balls	PB and CB	M6A	2	2/2	Open Response – Human Coded
M919Q01	CM919Q01S	PM919Q01	Fan Merchandise	PB and CB	M6A	3	1/2	Open Response – Computer Scored
M919Q02	CM919Q02S	PM919Q02	Fan Merchandise	PB and CB	M6A	4	2/2	Open Response – Computer Scored
M954Q01	CM954Q01S	PM954Q01	Medicine doses	PB and CB	M6A	5	1/3	Open Response – Computer Scored
M954Q02	DM954Q02C	PM954Q02	Medicine doses	PB and CB	M6A	6	2/3	Open Response – Human Coded
M954Q04	CM954Q04S	PM954Q04	Medicine doses	PB and CB	M6A	7	3/3	Open Response – Computer Scored
M943Q01	CM943Q01S	PM943Q01S	Arches	PB and CB	M6A	8	1/2	Simple Multiple Choice – Computer Scored
M943Q02	CM943Q02S	PM943Q02	Arches	PB and CB	M6A	9	2/2	Open Response – Computer Scored
M953Q02	DM953Q02C	PM953Q02	Flu test	PB and CB	M6A	10	1/3	Open Response – Human Coded
M953Q03	CM953Q03S	PM953Q03	Flu test	PB and CB	M6A	11	2/3	Open Response – Computer Scored
M953Q04	DM953Q04C	PM953Q04	Flu test	PB and CB	M6A	12	3/3	Open Response – Human Coded
M948Q01	CM948Q01S	PM948Q01S	Part Time Work	PB and CB	M6B	1	1/3	Simple Multiple Choice – Computer Scored
M948Q02	CM948Q02S	PM948Q02	Part Time Work	PB and CB	M6B	2	2/3	Open Response – Computer Scored
M948Q03	CM948Q03S	PM948Q03	Part Time Work	PB and CB	M6B	3	3/3	Open Response – Computer Scored
M936Q01	CM936Q01S	PM936Q01	Seats in a Theatre	PB and CB	M6B	4	1/2	Open Response – Computer Scored
M936Q02	DM936Q02C	PM936Q02	Seats in a Theatre	PB and CB	M6B	5	2/2	Open Response – Human Coded
M961Q02	DM961Q02C	PM961Q02	Chocolate	PB and CB	M6B	6	1/3	Open Response – Human Coded
M961Q03	CM961Q03S	PM961Q03S	Chocolate	PB and CB	M6B	7	2/3	Simple Multiple Choice – Computer Scored
M961Q05	DM961Q05C	PM961Q05	Chocolate	PB and CB	M6B	8	3/3	Open Response – Human Coded
M939Q01	CM939Q01S	PM939Q01S	Racing	PB and CB	M6B	9	1/2	Simple Multiple Choice – Computer Scored
M939Q02	CM939Q02S	PM939Q02S	Racing	PB and CB	M6B	10	2/2	Simple Multiple Choice – Computer Scored
M967Q01	CM967Q01S	PM967Q01	Wooden Train Set	PB and CB	M6B	11	1/2	Open Response – Computer Scored
M967Q03	CM967Q03S	PM967Q03S	Wooden Train Set	PB and CB	M6B	12	2/2	Complex Multiple Choice – Computer Scored
M192Q01	NA	PM192Q01S	Containers	PB	M2	5	1/1	NA
M192Q01	NA	PM828Q01	Carbon Dioxide	PB	M3	8	—	NA

## [Part 3/8]

**Table A.4 PISA 2015 main survey trend math item classification**

Generic ID	Item format – PBA	Content	Situation/context	Process	Unit origin	Language of submission	Version used as source for 2015
M033Q01	Simple Multiple Choice – Data Entered	Space and Shape	Personal	Interpreting, Applying and Evaluating Mathematical Outcomes	Consortium	Dutch	2012
M474Q01	Open Response – Human Coded	Quantity	Personal	Employing Mathematical Concepts, Facts and Procedures	Canada	English	2012
M155Q02	Open Response – Human Coded	Change and Relationships	Scientific	Interpreting, Applying and Evaluating Mathematical Outcomes	Consortium	Dutch	2012
M155Q01	Open Response – Human Coded	Change and Relationships	Scientific	Employing Mathematical Concepts, Facts and Procedures	Consortium	Dutch	2012
M155Q03	Open Response – Human Coded	Change and Relationships	Scientific	Employing Mathematical Concepts, Facts and Procedures	Consortium	Dutch	2012
M155Q04	Complex Multiple Choice – Data Entered	Change and Relationships	Scientific	Interpreting, Applying and Evaluating Mathematical Outcomes	Consortium	Dutch	2012
M411Q01	Open Response – Human Coded	Quantity	Societal	Employing Mathematical Concepts, Facts and Procedures	Consortium	English	2012
M411Q02	Simple Multiple Choice – Data Entered	Uncertainty and Data	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	Consortium	English	2012
M803Q01	Open Response – Data Entered	Uncertainty and Data	Occupational	Formulating Situations Mathematically	Canada	English	2012
M442Q02	Open Response – Human Coded	Quantity	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	Canada	English	2012
M462Q01	Open Response – Human Coded	Space and Shape	Scientific	Employing Mathematical Concepts, Facts and Procedures	Sweden	English	2012
M034Q01	Open Response – Data Entered	Space and Shape	Occupational	Formulating Situations Mathematically	Consortium	Dutch	2012
M305Q01	Simple Multiple Choice – Data Entered	Space and Shape	Societal	Employing Mathematical Concepts, Facts and Procedures			2012
M496Q01	Complex Multiple Choice – Data Entered	Quantity	Societal	Formulating Situations Mathematically	Consortium	English	2012
M496Q02	Open Response – Human Coded	Quantity	Societal	Employing Mathematical Concepts, Facts and Procedures	Consortium	English	2012
M423Q01	Simple Multiple Choice – Data Entered	Uncertainty and Data	Personal	Interpreting, Applying and Evaluating Mathematical Outcomes	Consortium	English	2012
M406Q01	Open Response – Human Coded	Space and Shape	Societal	Employing Mathematical Concepts, Facts and Procedures	Consortium	English	2012
M406Q02	Open Response – Human Coded	Space and Shape	Societal	Formulating Situations Mathematically	Consortium	English	2012
M603Q01	Complex Multiple Choice – Data Entered	Quantity	Scientific	Employing Mathematical Concepts, Facts and Procedures	Austria	German	2012
M571Q01	Simple Multiple Choice – Data Entered	Change and Relationships	Scientific	Interpreting, Applying and Evaluating Mathematical Outcomes	Germany	German	2012
M564Q01	Simple Multiple Choice – Data Entered	Quantity	Societal	Formulating Situations Mathematically	Italy	English	2012
M564Q02	Simple Multiple Choice – Data Entered	Uncertainty and Data	Societal	Formulating Situations Mathematically	Italy	English	2012
M447Q01	Simple Multiple Choice – Data Entered	Space and Shape	Societal	Employing Mathematical Concepts, Facts and Procedures	Consortium	English	2012
M273Q01	Complex Multiple Choice – Data Entered	Space and Shape	Occupational	Employing Mathematical Concepts, Facts and Procedures	Czech Republic	Czech	2012
R111Q06	Complex Multiple Choice – Data Entered	Uncertainty and Data	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	Consortium	English	2012
M420Q01	Complex Multiple Choice – Data Entered	Uncertainty and Data	Personal	Interpreting, Applying and Evaluating Mathematical Outcomes	Consortium	English	2012
M446Q01	Open Response – Human Coded	Change and Relationships	Scientific	Formulating Situations Mathematically	Consortium	English	2012
M446Q02	Open Response – Human Coded	Change and Relationships	Scientific	Formulating Situations Mathematically	Consortium	English	2012
M559Q01	Simple Multiple Choice – Data Entered	Quantity	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	Italy	English	2012
M828Q02	Open Response – Human Coded	Uncertainty and Data	Scientific	Employing Mathematical Concepts, Facts and Procedures	Netherlands	English	2012
M828Q03	Open Response – Human Coded	Quantity	Scientific	Employing Mathematical Concepts, Facts and Procedures	Netherlands	English	2012
M464Q01	Open Response – Data Entered	Space and Shape	Societal	Formulating Situations Mathematically	Sweden	English	2012
M800Q01	Simple Multiple Choice – Data Entered	Quantity	Personal	Employing Mathematical Concepts, Facts and Procedures	Canada	English	2012
M982Q01	Open Response – Human Coded	Uncertainty and Data	Societal	Employing Mathematical Concepts, Facts and Procedures	ACER	English	2012
M982Q02	Open Response – Human Coded	Uncertainty and Data	Societal	Employing Mathematical Concepts, Facts and Procedures	ACER	English	2012
M982Q03	Complex Multiple Choice – Data Entered	Uncertainty and Data	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	ACER	English	2012
M982Q04	Simple Multiple Choice – Data Entered	Uncertainty and Data	Societal	Formulating Situations Mathematically	ACER	English	2012
M992Q01	Open Response – Human Coded	Space and Shape	Occupational	Formulating Situations Mathematically	France	English	2012
M992Q02	Open Response – Human Coded	Space and Shape	Occupational	Formulating Situations Mathematically	France	English	2012
M992Q03	Open Response – Human Coded	Change and Relationships	Occupational	Formulating Situations Mathematically	France	English	2012
M915Q01	Simple Multiple Choice – Data Entered	Uncertainty and Data	Societal	Employing Mathematical Concepts, Facts and Procedures	ILS	English	2012
M915Q02	Open Response – Human Coded	Change and Relationships	Societal	Employing Mathematical Concepts, Facts and Procedures	ILS	English	2012



## [Part 4/8]

Table A.4 PISA 2015 main survey trend math item classification

Generic ID	Item format – PBA	Content	Situation/context	Process	Unit origin	Language of submission	Version used as source for 2015
M906Q01	Simple Multiple Choice – Data Entered	Quantity	Scientific	Employing Mathematical Concepts, Facts and Procedures	ACER	English	2012
M906Q02	Open Response – Human Coded	Quantity	Scientific	Employing Mathematical Concepts, Facts and Procedures	ACER	English	2012
M00KQ02	Open Response – Human Coded	Space and Shape	Personal	Formulating Situations Mathematically	Canada	English	2012
M909Q01	Open Response – Human Coded	Quantity	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	aSPe	English	2012
M909Q02	Simple Multiple Choice – Data Entered	Quantity	Societal	Employing Mathematical Concepts, Facts and Procedures	aSPe	English	2012
M909Q03	Open Response – Human Coded	Change and Relationships	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	aSPe	English	2012
M949Q01	Complex Multiple Choice – Data Entered	Space and Shape	Occupational	Employing Mathematical Concepts, Facts and Procedures	ACER	English	2012
M949Q02	Complex Multiple Choice – Data Entered	Space and Shape	Occupational	Employing Mathematical Concepts, Facts and Procedures	ACER	English	2012
M949Q03	Open Response – Human Coded	Space and Shape	Occupational	Formulating Situations Mathematically	ACER	English	2012
M00GQ01	Open Response – Human Coded	Space and Shape	Personal	Formulating Situations Mathematically	Czech Republic	Czech	2012
M955Q01	Open Response – Human Coded	Uncertainty and Data	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	University of Melbourne	English	2012
M955Q02	Open Response – Human Coded	Uncertainty and Data	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	University of Melbourne	English	2012
M955Q03	Open Response – Human Coded	Uncertainty and Data	Societal	Employing Mathematical Concepts, Facts and Procedures	University of Melbourne	English	2012
M998Q02	Open Response – Human Coded	Change and Relationships	Personal	Interpreting, Applying and Evaluating Mathematical Outcomes	Israel	English	2012
M998Q04	Complex Multiple Choice – Data Entered	Change and Relationships	Personal	Employing Mathematical Concepts, Facts and Procedures	Israel	English	2012
M905Q01	Complex Multiple Choice – Data Entered	Quantity	Occupational	Interpreting, Applying and Evaluating Mathematical Outcomes	ACER	English	2012
M905Q02	Open Response – Human Coded	Quantity	Occupational	Interpreting, Applying and Evaluating Mathematical Outcomes	ACER	English	2012
M919Q01	Open Response – Human Coded	Quantity	Personal	Employing Mathematical Concepts, Facts and Procedures	IPN/Kassel	English	2012
M919Q02	Open Response – Human Coded	Quantity	Personal	Formulating Situations Mathematically	IPN/Kassel	English	2012
M954Q01	Open Response – Human Coded	Change and Relationships	Scientific	Employing Mathematical Concepts, Facts and Procedures	University of Melbourne	English	2012
M954Q02	Open Response – Human Coded	Change and Relationships	Scientific	Employing Mathematical Concepts, Facts and Procedures	University of Melbourne	English	2012
M954Q04	Open Response – Human Coded	Change and Relationships	Scientific	Employing Mathematical Concepts, Facts and Procedures	University of Melbourne	English	2012
M943Q01	Simple Multiple Choice – Data Entered	Change and Relationships	Occupational	Formulating Situations Mathematically	ACER	English	2012
M943Q02	Open Response – Human Coded	Space and Shape	Occupational	Formulating Situations Mathematically	ACER	English	2012
M953Q02	Open Response – Human Coded	Uncertainty and Data	Scientific	Interpreting, Applying and Evaluating Mathematical Outcomes	University of Melbourne	English	2012
M953Q03	Open Response – Human Coded	Uncertainty and Data	Scientific	Formulating Situations Mathematically	University of Melbourne	English	2012
M953Q04	Open Response – Human Coded	Uncertainty and Data	Scientific	Formulating Situations Mathematically	University of Melbourne	English	2012
M948Q01	Simple Multiple Choice – Data Entered	Quantity	Occupational	Interpreting, Applying and Evaluating Mathematical Outcomes	ACER	English	2012
M948Q02	Open Response – Human Coded	Quantity	Occupational	Employing Mathematical Concepts, Facts and Procedures	ACER	English	2012
M948Q03	Open Response – Human Coded	Quantity	Occupational	Employing Mathematical Concepts, Facts and Procedures	ACER	English	2012
M936Q01	Open Response – Human Coded	Change and Relationships	Occupational	Employing Mathematical Concepts, Facts and Procedures	MEG	English	2012
M936Q02	Open Response – Human Coded	Change and Relationships	Occupational	Formulating Situations Mathematically	MEG	English	2012
M961Q02	Open Response – Human Coded	Change and Relationships	Occupational	Employing Mathematical Concepts, Facts and Procedures	IPN/Kassel	English	2012
M961Q03	Simple Multiple Choice – Data Entered	Change and Relationships	Scientific	Employing Mathematical Concepts, Facts and Procedures	IPN/Kassel	English	2012
M961Q05	Open Response – Human Coded	Uncertainty and Data	Occupational	Interpreting, Applying and Evaluating Mathematical Outcomes	IPN/Kassel	English	2012
M939Q01	Simple Multiple Choice – Data Entered	Uncertainty and Data	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	MEG	English	2012
M939Q02	Simple Multiple Choice – Data Entered	Uncertainty and Data	Societal	Interpreting, Applying and Evaluating Mathematical Outcomes	MEG	English	2012
M967Q01	Open Response – Human Coded	Space and Shape	Personal	Employing Mathematical Concepts, Facts and Procedures	IPN/Kassel	English	2012
M967Q03	Complex Multiple Choice – Data Entered	Space and Shape	Personal	Formulating Situations Mathematically	IPN/Kassel	English	2012
M192Q01	Complex Multiple Choice – Data Entered	Change and Relationships	Scientific	Formulating Situations Mathematically	Germany	German	2012
M192Q01	Open Response – Human Coded	Change and Relationships	Scientific	Employing Mathematical Concepts, Facts and Procedures	Netherlands	English	2012

[Part 5/8]

Table A.4 PISA 2015 main survey trend math item classification

Generic ID	CBA international % correct	CBA international % correct S.E.	CBA item parameters (RP = 0.50)				CBA thresholds (RP = 0.62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
M033Q01	74.66	0.22	1.000	-0.956			423		Level 2
M474Q01	62.20	0.24	1.000	-0.670			462		Level 2
M155Q02	54.57	0.22	1.000	-0.357	-0.424	0.424	466	510	Level 3
M155Q01	62.23	0.25	1.000	-0.672			462		Level 2
M155Q03	17.44	0.18	1.087	0.735	-0.201	0.201	614	663	Level 5
M155Q04	48.01	0.25	1.000	-0.276			516		Level 3
M411Q01	41.46	0.25	1.426	-0.075			531		Level 3
M411Q02	40.97	0.25	0.686	0.055			579		Level 4
M803Q01	20.86	0.21	1.695	0.514			607		Level 5
M442Q02	24.37	0.22	1.483	0.438			600		Level 4
M462Q01	9.68	0.14	0.883	1.142	-0.374	0.374	669	724	Level 6
M034Q01	33.77	0.25	1.000	0.251			587		Level 4
M305Q01	40.34	0.25	0.621	0.149			597		Level 4
M496Q01	41.85	0.25	1.000	-0.023			550		Level 4
M496Q02	60.48	0.25	1.000	-0.610			470		Level 2
M423Q01	75.91	0.22	0.586	-1.548			371		Level 1
M406Q01	20.64	0.21	1.780	0.512			606		Level 4
M406Q02	10.94	0.16	2.303	0.783			638		Level 5
M603Q01	33.78	0.24	0.765	0.326			610		Level 5
M571Q01	39.84	0.26	1.000	0.031			558		Level 4
M564Q01	46.10	0.26	0.631	-0.042			571		Level 4
M564Q02	44.38	0.26	1.000	-0.094			541		Level 3
M447Q01	58.66	0.24	1.000	-0.545			479		Level 2
M273Q01	41.32	0.25	0.737	0.061			576		Level 4
R111Q06	32.15	0.23	1.056	0.305			593		Level 4
M420Q01	43.22	0.25	0.840	-0.044			555		Level 4
M446Q01	59.88	0.25	1.402	-0.561			466		Level 2
M446Q02	7.09	0.13	1.000	1.705			785		Level 6
M559Q01	58.74	0.25	1.000	-0.616			470		Level 2
M828Q02	57.80	0.25	1.006	-0.534			480		Level 2
M828Q03	27.75	0.23	1.070	0.446			611		Level 5
M464Q01	20.59	0.21	1.643	0.523			609		Level 5
M800Q01	88.52	0.16	1.000	-1.805			308		Below Level 1
M982Q01	81.34	0.20	1.000	-1.475			353		Below Level 1
M982Q02	29.39	0.23	0.830	0.511			631		Level 5
M982Q03	61.83	0.25	1.000	-0.629			468		Level 2
M982Q04	43.61	0.25	1.087	-0.055			543		Level 3
M992Q01	70.40	0.23	1.000	-0.936			426		Level 2
M992Q02	14.53	0.18	1.321	0.883			664		Level 5
M992Q03	6.85	0.14	2.097	1.018			671		Level 6
M915Q01	38.50	0.25	0.830	0.152			582		Level 4
M915Q02	63.40	0.25	1.232	-0.737			446		Level 2



[Part 6/8]

Table A.4 PISA 2015 main survey trend math item classification

Generic ID	CBA international % correct	CBA international % correct S.E.	CBA item parameters (RP = 0.50)				CBA thresholds (RP = 0.62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
M906Q01	57.25	0.25	1.000	-0.499			486		Level 3
M906Q02	39.53	0.23	1.011	-0.009	-0.541	0.541	513	552	Level 4
M00KQ02	12.45	0.18	1.000	1.116			705		Level 6
M909Q01	81.00	0.21	1.000	-1.455			356		Below Level 1
M909Q02	52.93	0.25	1.000	-0.442			493		Level 3
M909Q03	27.40	0.23	1.821	0.258			571		Level 4
M949Q01	62.51	0.25	1.282	-0.642			457		Level 2
M949Q02	31.91	0.23	1.229	0.165			568		Level 4
M949Q03	28.09	0.23	0.631	0.313	-2.368	2.368	557	594	Level 4
M00GQ01	6.19	0.13	1.000	1.622			774		Level 6
M955Q01	65.19	0.23	0.857	-0.850			444		Level 2
M955Q02	29.96	0.24	1.274	0.308			587		Level 4
M955Q03	9.31	0.14	1.000	0.990	-0.773	0.773	649	682	Level 6
M998Q02	65.30	0.24	0.996	-0.734			454		Level 2
M998Q04	37.26	0.28	0.220	1.132			846		Level 6
M905Q01	74.44	0.25	1.000	-0.937			426		Level 2
M905Q02	39.16	0.29	1.870	0.122			552		Level 4
M919Q01	80.97	0.22	1.000	-1.246			384		Level 1
M919Q02	43.07	0.29	0.830	0.100			575		Level 4
M954Q01	66.99	0.28	1.505	-0.611			457		Level 2
M954Q02	33.15	0.27	1.286	0.361			594		Level 4
M954Q04	25.38	0.25	1.591	0.500			607		Level 5
M943Q01	52.95	0.29	0.737	-0.202			540		Level 3
M943Q02	5.26	0.14	1.804	1.331			717		Level 6
M953Q02	42.17	0.30	1.000	0.084			565		Level 4
M953Q03	53.34	0.31	1.662	-0.189			512		Level 3
M953Q04	19.87	0.25	1.000	0.730	-0.427	0.427	613	657	Level 5
M948Q01	77.19	0.44	1.141	-1.702			317		Below Level 1
M948Q02	53.57	0.54	1.000	-1.015			415		Level 1
M948Q03	6.05	0.26	1.000	0.978			686		Level 6
M936Q01	33.25	0.54	1.934	-0.462			472		Level 2
M936Q02	26.13	0.47	1.855	-0.399			481		Level 2
M961Q02	3.98	0.21	1.000	1.287			728		Level 6
M961Q03	35.84	0.53	1.000	-0.471			489		Level 3
M961Q05	34.61	0.47	0.705	-0.410	-0.190	0.190	459	542	Level 3
M939Q01	47.78	0.54	0.499	-0.560			517		Level 3
M939Q02	34.61	0.52	1.000	-0.312			511		Level 3
M967Q01	22.80	0.47	1.593	-0.163			517		Level 3
M967Q03	6.22	0.27	1.000	0.960			684		Level 6
M192Q01	NA	NA	NA	NA	NA	NA	NA	NA	NA
M192Q01	NA	NA	NA	NA	NA	NA	NA	NA	NA

## [Part 7/8]

Table A.4 PISA 2015 main survey trend math item classification

Generic ID	PBA international % correct	PBA international % correct S.E.	PBA item parameters (RP = 0.50)				PBA thresholds (RP = 0.62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
M033Q01	51.33	0.53	1.000	-0.956			423		Level 2
M474Q01	49.13	0.54	1.000	-0.842			439		Level 2
M155Q02	29.15	0.47	1.000	-0.357	-0.424	0.424	466	510	Level 3
M155Q01	37.49	0.53	1.000	-0.582			474		Level 2
M155Q03	6.52	0.24	1.087	0.735	-0.201	0.201	614	663	Level 5
M155Q04	34.77	0.47	1.000	-0.276			516		Level 3
M411Q01	23.70	0.48	1.426	-0.075			531		Level 3
M411Q02	32.33	0.47	0.686	0.055			579		Level 4
M803Q01	8.41	0.32	1.695	0.444			598		Level 4
M442Q02	16.73	0.42	1.483	0.171			564		Level 4
M462Q01	6.71	0.28	0.883	1.142	-0.374	0.374	669	724	Level 6
M034Q01	20.40	0.49	1.000	0.158			575		Level 4
M305Q01	37.97	0.53	0.621	-0.534			505		Level 3
M496Q01	30.99	0.49	1.000	-0.203			526		Level 3
M496Q02	39.52	0.53	1.000	-0.608			471		Level 2
M423Q01	62.99	0.52	0.586	-1.548			371		Level 1
M406Q01	9.54	0.32	1.780	0.419			593		Level 4
M406Q02	4.92	0.24	2.303	0.664			621		Level 5
M603Q01	30.29	0.49	0.765	0.085			577		Level 4
M571Q01	29.96	0.49	1.000	-0.048			547		Level 4
M564Q01	35.51	0.52	0.631	-0.042			571		Level 4
M564Q02	36.80	0.49	1.000	-0.094			541		Level 3
M447Q01	40.52	0.53	1.000	-0.667			463		Level 2
M273Q01	36.97	0.51	0.737	-0.257			532		Level 3
R111Q06	17.75	0.41	1.056	0.210			580		Level 4
M420Q01	31.04	0.46	0.840	-0.044			555		Level 4
M446Q01	36.88	0.52	1.402	-0.561			466		Level 2
M446Q02	2.79	0.18	1.000	1.705			785		Level 6
M559Q01	48.88	0.57	1.000	-0.616			470		Level 2
M828Q02	33.05	0.50	1.006	-0.319			510		Level 3
M828Q03	15.13	0.40	1.070	0.446			611		Level 5
M464Q01	8.98	0.35	1.643	0.523			609		Level 5
M800Q01	84.06	0.42	1.000	-1.805			308		Below Level 1
M982Q01	73.49	0.49	1.000	-1.556			342		Below Level 1
M982Q02	20.81	0.44	0.830	0.511			631		Level 5
M982Q03	50.66	0.53	1.000	-0.672			462		Level 2
M982Q04	29.52	0.46	1.087	-0.123			533		Level 3
M992Q01	60.00	0.54	1.000	-1.096			404		Level 1
M992Q02	11.05	0.35	1.321	0.743			645		Level 5
M992Q03	4.78	0.24	2.097	1.018			671		Level 6
M915Q01	25.87	0.47	0.830	0.152			582		Level 4
M915Q02	52.61	0.56	1.232	-0.737			446		Level 2



[Part 8/8]

Table A.4 PISA 2015 main survey trend math item classification

Generic ID	PBA international % correct	PBA international % correct S.E.	PBA item parameters (RP = 0.50)				PBA thresholds (RP = 0.62)		Level
			Slope	Difficulty	Step 1	Step 2	1.00	2.00	
M906Q01	45.19	0.54	1.000	-0.499			486		Level 3
M906Q02	26.59	0.46	1.011	-0.009	-0.541	0.541	513	552	Level 4
M00KQ02	11.24	0.35	1.000	1.116			705		Level 6
M909Q01	67.65	0.49	1.000	-1.607			335		Below Level 1
M909Q02	39.67	0.50	1.000	-0.509			484		Level 3
M909Q03	12.03	0.36	1.821	0.202			563		Level 4
M949Q01	49.61	0.53	1.282	-0.642			457		Level 2
M949Q02	27.37	0.51	1.229	0.165			568		Level 4
M949Q03	20.71	0.45	0.631	0.313	-2.368	2.368	557	594	Level 4
M00GQ01	3.60	0.20	1.000	1.622			774		Level 6
M955Q01	50.85	0.53	0.857	-0.850			444		Level 2
M955Q02	24.29	0.47	1.274	0.308			587		Level 4
M955Q03	5.38	0.23	1.000	0.977	-0.609	0.609	647	685	Level 6
M998Q02	42.56	0.56	0.996	-0.734			454		Level 2
M998Q04	37.24	0.55	0.220	1.132			846		Level 6
M905Q01	56.99	0.98	1.000	-1.019			415		Level 1
M905Q02	30.17	0.90	1.870	-0.112			520		Level 3
M919Q01	57.50	1.02	1.000	-1.339			371		Level 1
M919Q02	24.55	0.95	0.830	0.100			575		Level 4
M954Q01	48.87	1.00	1.505	-0.611			457		Level 2
M954Q02	14.42	0.69	1.286	0.361			594		Level 4
M954Q04	14.72	0.62	1.591	0.500			607		Level 5
M943Q01	41.91	0.99	0.737	-0.202			540		Level 3
M943Q02	4.59	0.48	1.804	1.331			717		Level 6
M953Q02	31.63	0.88	1.000	-0.084			542		Level 3
M953Q03	35.01	1.01	1.662	-0.142			518		Level 3
M953Q04	12.66	0.77	1.000	0.762	-0.501	0.501	618	659	Level 5
M948Q01	70.10	0.56	1.141	-1.702			317		Below Level 1
M948Q02	57.14	0.64	1.000	-1.015			415		Level 1
M948Q03	9.35	0.38	1.000	0.978			686		Level 6
M936Q01	40.66	0.72	1.934	-0.462			472		Level 2
M936Q02	34.60	0.64	1.855	-0.399			481		Level 2
M961Q02	8.62	0.34	1.000	1.287			728		Level 6
M961Q03	40.04	0.61	1.000	-0.471			489		Level 3
M961Q05	37.47	0.56	0.705	-0.410	-0.190	0.190	459	542	Level 3
M939Q01	49.65	0.67	0.499	-0.560			517		Level 3
M939Q02	37.92	0.64	1.000	-0.312			511		Level 3
M967Q01	28.94	0.60	1.593	-0.163			517		Level 3
M967Q03	7.01	0.37	1.000	0.960			684		Level 6
M192Q01	18.85	0.43	1.000	0.209			582		Level 4
M192Q01	17.04	0.39	1.372	0.256			577		Level 4

## [Part 1/3]

Table A.5 PISA 2015 main survey financial literacy item classification

Generic ID	CBA item ID in main survey analysis output	Unit ID	Mode [paper-based (PB); computer-based (CB)]	2015 field trial and main survey cluster	Sequence in cluster	Sequence in unit	Item format
F001Q01	CF001Q01S	Costs of Running a Car	CB	F1	3	1/1	Complex Multiple Choice – Computer Scored
F004Q03	DF004Q03C	Income tax	CB	F2	20	1/1	Open Response – Human Coded
F006Q02	CF006Q02S	Music system	CB	F2	4	1/1	Complex Multiple Choice – Computer Scored
F009Q02	CF009Q02S	Shopping	CB	F1	1	1/1	Simple Multiple Choice – Computer Scored
F010Q01	CF010Q01S	Bank statement	CB	F1	14	1/2	Open Response – Computer Scored
F010Q02	CF010Q02S	Bank statement	CB	F1	15	2/2	Open Response – Computer Scored
F012Q01	CF012Q01S	Interest	CB	F1	12	1/2	Simple Multiple Choice – Computer Scored
F012Q02	CF012Q02S	Interest	CB	F1	13	2/2	Complex Multiple Choice – Computer Scored
F024Q02	DF024Q02C	Jacket sale	CB	F2	12	1/1	Open Response – Human Coded
F028Q02	DF028Q02C	Phone plans	CB	F1	5	1/2	Open Response – Human Coded
F028Q03	CF028Q03S	Phone plans	CB	F1	6	2/2	Simple Multiple Choice – Computer Scored
F031Q01	CF031Q01S	Laptop	CB	F1	10	1/2	Complex Multiple Choice – Computer Scored
F031Q02	CF031Q02S	Laptop	CB	F1	11	2/2	Open Response – Computer Scored
F033Q01	CF033Q01S	Wayne's Bank Statement	CB	F2	13	1/2	Simple Multiple Choice – Computer Scored
F033Q02	CF033Q02S	Wayne's Bank Statement	CB	F2	14	2/2	Simple Multiple Choice – Computer Scored
F035Q01	CF035Q01S	Ring-Tones	CB	F2	16	1/1	Open Response – Computer Scored
F036Q01	DF036Q01C	Online Shopping	CB	F1	17	1/1	Open Response – Human Coded
F051Q01	DF051Q01C	Bicycle Shop	CB	F2	6	1/2	Open Response – Human Coded
F051Q02	DF051Q02C	Bicycle Shop	CB	F2	7	2/2	Open Response – Human Coded
F052Q01	CF052Q01S	Video Game	CB	F2	9	1/1	Complex Multiple Choice – Computer Scored
F054Q01	DF054Q01C	E-mail	CB	F1	4	1/1	Open Response – Human Coded
F058Q01	DF058Q01C	PIN	CB	F2	3	1/1	Open Response – Human Coded
F062Q01	CF062Q01S	Mobile Phone Contract	CB	F2	8	1/1	Complex Multiple Choice – Computer Scored
F068Q01	DF068Q01C	Job Change	CB	F1	9	1/1	Open Response – Human Coded
F069Q01	CF069Q01S	Student Account	CB	F2	5	1/1	Complex Multiple Choice – Computer Scored
F075Q02	CF075Q02S	Study Options	CB	F2	17	1/1	Complex Multiple Choice – Computer Scored
F082Q01	DF082Q01C	New Bike	CB	F1	7	1/2	Open Response – Human Coded
F082Q02	CF082Q02S	New Bike	CB	F1	8	2/2	Simple Multiple Choice – Computer Scored
F095Q01	CF095Q01S	Changing Value	CB	F2	18	1/2	Simple Multiple Choice – Computer Scored
F095Q02	CF095Q02S	Changing Value	CB	F2	19	2/2	Complex Multiple Choice – Computer Scored
F097Q01	CF097Q01S	Company Profit	CB	F1	19	1/1	Complex Multiple Choice – Computer Scored
F102Q01	CF102Q01S	Gantica	CB	F2	1	1/2	Open Response – Computer Scored
F102Q02	DF102Q02C	Gantica	CB	F2	2	2/2	Open Response – Human Coded
F103Q01	DF103Q01C	Investing	CB	F1	18	1/1	Open Response – Human Coded
F105Q01	CF105Q01S	Interest Rates	CB	F1	21	1/2	Simple Multiple Choice – Computer Scored
F105Q02	CF105Q02S	Interest Rates	CB	F1	22	2/2	Simple Multiple Choice – Computer Scored
F106Q01	DF106Q01C	Family Holiday	CB	F2	10	1/2	Open Response – Human Coded
F106Q02	CF106Q02S	Family Holiday	CB	F2	11	2/2	Simple Multiple Choice – Computer Scored
F110Q01	CF110Q01S	Living Alone	CB	F1	2	1/1	Complex Multiple Choice – Computer Scored
F200Q01	DF200Q01C	Charitable Giving	CB	F1	20	1/1	Open Response – Human Coded
F201Q01	DF201Q01C	Emergency Funds	CB	F1	16	1/1	Open Response – Human Coded
F202Q01	CF202Q01S	Book Purchase	CB	F2	15	1/1	Complex Multiple Choice – Computer Scored
F203Q01	DF203Q01C	No Credit	CB	F2	21	1/1	Open Response – Human Coded



[Part 2/3]

Table A.5 PISA 2015 main survey financial literacy item classification

Generic ID	Content	Process	Context	Unit origin	Language of submission	Source
F001Q01	Planning and Managing Finances	Analyse information in a financial context	Home and family	ACER	English	2012
F004Q03	Planning and Managing Finances	Evaluate financial Issues	Education and work	ACER	English	2012
F006Q02	Planning and Managing Finances	Analyse information in a financial context	Individual	ACER	English	2012
F009Q02	Money and Transactions	Apply financial knowledge and understanding	Home and family	ACER	English	2012
F010Q01	Money and Transactions	Identify financial information	Home and family	ACER	English	2012
F010Q02	Money and Transactions	Analyse information in a financial context	Home and family	ACER	English	2012
F012Q01	Risk and reward	Apply financial knowledge and understanding	Individual	ACER	English	2012
F012Q02	Risk and Reward	Analyse information in a financial context	Individual	ACER	English	2012
F024Q02	Money and Transactions	Evaluate financial issues	Individual	ACER	English	2012
F028Q02	Planning and Managing Finances	Analyse information in a financial context	Individual	ACER	English	2012
F028Q03	Planning and Managing Finances	Analyse information in a financial context	Individual	ACER	English	2012
F031Q01	Risk and Reward	Evaluate financial issues	Home and family	ACER	English	2012
F031Q02	Risk and reward	Apply financial knowledge and understanding	Home and family	ACER	English	2012
F033Q01	Money and Transactions	Analyse information in a financial context	Individual	ACER	English	2012
F033Q02	Money and Transactions	Identify financial information	Individual	ACER	English	2012
F035Q01	Financial Landscape	Apply financial knowledge and understanding	Individual	ACER	English	2012
F036Q01	Financial Landscape	Evaluate financial issues	Societal	ACER	English	2012
F051Q01	Planning and Managing Finances	Evaluate financial issues	Education and work	ACER	English	2012
F051Q02	Planning and Managing Finances	Evaluate financial issues	Education and work	ACER	English	2012
F052Q01	Planning and Managing Finances	Identify financial information	Individual	ACER	English	2012
F054Q01	Financial Landscape	Evaluate financial issues	Societal	ACER	English	2012
F058Q01	Risk and Reward	Evaluate financial issues	Societal	ACER	English	2012
F062Q01	Financial Landscape	Evaluate financial issues	Home and family	ACER	English	2012
F068Q01	Planning and Managing Finances	Evaluate financial issues	Education and work	ACER	English	2012
F069Q01	Financial Landscape	Analyse information in a financial context	Education and work	ACER	English	2012
F075Q02	Planning and Managing Finances	Analyse information in a financial context	Education and work	ACER	English	2012
F082Q01	Money and Transactions	Identify financial information	Individual	ACER	English	2012
F082Q02	Risk and Reward	Identify financial information	Home and family	ACER	English	2012
F095Q01	Money and Transactions	Identify financial information	Home and family	ACER	English	2012
F095Q02	Financial Landscape	Analyse information in a financial context	Societal	ACER	English	2012
F097Q01	Financial Landscape	Identify financial information	Individual	ACER	English	2012
F102Q01	Risk and Reward	Apply financial knowledge and understanding	Home and family	ACER	English	2012
F102Q02	Risk and Reward	Apply financial knowledge and understanding	Home and family	ACER	English	2012
F103Q01	Risk and Reward	Evaluate financial issues	Individual	ACER	English	2012
F105Q01	Money and Transactions	Apply financial knowledge and understanding	Individual	ACER	English	2012
F105Q02	Money and Transactions	Apply financial knowledge and understanding	Individual	ACER	English	2012
F106Q01	Planning and Managing Finances	Evaluate financial issues	Home and family	ACER	English	2012
F106Q02	Planning and Managing Finances	Apply financial knowledge and understanding	Home and family	ACER	English	2012
F110Q01	Planning and Managing Finances	Evaluate financial issues	Home and family	ACER	English	2012
F200Q01	Financial Landscape	Evaluate financial issues	Societal	ETS	English	2015
F201Q01	Planning and Managing Finances	Analyse information in a financial context	Individual	ETS	English	2015
F202Q01	Money and Transactions	Apply financial knowledge and understanding	Home and family	ETS	English	2015
F203Q01	Financial Landscape	Evaluate financial Issues	Individual	ETS	English	2015

[Part 3/3]

Table A.5 PISA 2015 main survey financial literacy item classification

Generic ID	International % correct	International % correct S.E.	Item parameters (RP = 0.50)				Thresholds (RP = .62)		Level
			Slope	Difficulty	Step 1	Step 2	1	2	
F001Q01	60.37	0.36	0.807	-0.400			485		Level 3
F004Q03	4.08	0.17	1.431	1.847			778		Level 5
F006Q02	43.78	0.35	0.664	0.227			583		Level 4
F009Q02	88.32	0.22	0.980	-1.644			302		Below Level 1
F010Q01	32.36	0.33	1.296	0.496			591		Level 4
F010Q02	16.62	0.24	1.066	0.988	-0.196	0.196	629	682	Level 5
F012Q01	53.43	0.35	0.658	-0.187			526		Level 3
F012Q02	48.45	0.35	0.110	0.421			917		Level 5
F024Q02	50.73	0.34	0.815	0.027			544		Level 3
F028Q02	43.33	0.36	1.058	0.108			544		Level 3
F028Q03	58.88	0.34	0.765	-0.373			491		Level 3
F031Q01	23.86	0.30	0.286	2.287			952		Level 5
F031Q02	48.30	0.34	0.866	0.068			547		Level 3
F033Q01	26.12	0.31	0.701	0.983			686		Level 5
F033Q02	52.69	0.35	0.606	-0.189			531		Level 3
F035Q01	30.12	0.33	0.996	0.610			617		Level 4
F036Q01	42.98	0.37	0.899	0.161			558		Level 4
F051Q01	74.46	0.34	1.305	-0.782			412		Level 2
F051Q02	36.00	0.35	0.925	0.398			590		Level 4
F052Q01	53.52	0.38	1.068	-0.108			513		Level 3
F054Q01	54.03	0.34	0.618	-0.221			525		Level 3
F058Q01	75.96	0.30	0.674	-1.459			346		Level 1
F062Q01	59.46	0.35	0.893	-0.465			471		Level 2
F068Q01	38.50	0.37	1.075	0.288			569		Level 4
F069Q01	58.28	0.36	1.149	-0.264			489		Level 3
F075Q02	28.39	0.30	0.329	1.727			855		Level 5
F082Q01	60.01	0.33	0.573	-0.376	-0.518	0.518	438	527	Level 3
F082Q02	76.09	0.32	0.732	-1.244			372		Level 1
F095Q01	37.57	0.35	0.634	0.577			635		Level 5
F095Q02	22.29	0.31	1.030	0.933			661		Level 5
F097Q01	7.85	0.21	1.349	1.519			733		Level 5
F102Q01	76.57	0.32	1.210	-0.917			396		Level 1
F102Q02	50.84	0.26	0.562	-0.028	0.635	-0.635	487	669	Level 5
F103Q01	25.65	0.32	1.346	0.708			620		Level 4
F105Q01	26.81	0.33	0.758	0.897			670		Level 5
F105Q02	35.97	0.35	0.898	0.424			595		Level 4
F106Q01	68.16	0.35	1.092	-0.692			431		Level 2
F106Q02	40.22	0.36	0.651	0.390			607		Level 4
F110Q01	79.50	0.29	1.133	-1.199			358		Level 1
F200Q01	58.12	0.33	0.874	-0.568			457		Level 2
F201Q01	57.40	0.37	1.136	-0.261			490		Level 3
F202Q01	38.55	0.36	0.776	0.420			602		Level 4
F203Q01	38.03	0.35	0.319	0.552	-3.862	3.862	568	652	Level 5



[Part 1/6]

Table A.6 PISA 2015 main survey CPS item classification

Item ID in analysis output	Unit name	C020C015 main survey cluster	Part	Score points	CPS skills	CPS competencies
CC104101	Meeting in the Park	C01	1	1	B2	Taking appropriate action to solve the problem
CC104102	Meeting in the Park	C01	1	1	A1	Establishing and maintaining shared understanding
CC104103	Meeting in the Park	C01	1	1	A2	Taking appropriate action to solve the problem
CC104105	Meeting in the Park	C01	1	1	B1	Establishing and maintaining shared understanding
CC104106	Meeting in the Park	C01	1	1	B1	Establishing and maintaining shared understanding
CC104107	Meeting in the Park	C01	1	1	B1	Establishing and maintaining shared understanding
CC106101	Making a Film	C01	1	1	A1	Establishing and maintaining shared understanding
CC106102	Making a Film	C01	1	1	A1	Establishing and maintaining shared understanding
CC106103	Making a Film	C01	1	1	A1	Establishing and maintaining shared understanding
CC106104	Making a Film	C01	1	1	C3	Establishing and maintaining team organisation
CC106105	Making a Film	C01	1	1	A2	Taking appropriate action to solve the problem
CC106106	Making a Film	C01	1	1	A1	Establishing and maintaining shared understanding
CC106107C	Making a Film	C01	1	3, 2, 1, 0	C3	Establishing and maintaining team organisation
CC104201	Meeting in the Park	C01	2	1	B3	Establishing and maintaining team organisation
CC104202	Meeting in the Park	C01	2	1	B1	Establishing and maintaining shared understanding
CC104203	Meeting in the Park	C01	2	1	B1	Establishing and maintaining shared understanding
CC104204	Meeting in the Park	C01	2	1	D3	Establishing and maintaining team organisation
CC104205	Meeting in the Park	C01	2	1	B2	Taking appropriate action to solve the problem
CC104206	Meeting in the Park	C01	2	1	C1	Establishing and maintaining shared understanding
CC106201	Making a Film	C01	2	1	A1	Establishing and maintaining shared understanding
CC106202	Making a Film	C01	2	1	B3	Establishing and maintaining team organisation
CC106203	Making a Film	C01	2	1	B3	Establishing and maintaining team organisation
CC106204	Making a Film	C01	2	1	B2	Taking appropriate action to solve the problem
CC106205	Making a Film	C01	2	1	B2	Taking appropriate action to solve the problem
CC106206	Making a Film	C01	2	1	D3	Establishing and maintaining team organisation
CC106207	Making a Film	C01	2	1	C3	Establishing and maintaining team organisation
CC106208	Making a Film	C01	2	1	C3	Establishing and maintaining team organisation
CC106209	Making a Film	C01	2	2, 1	C2	Taking appropriate action to solve the problem
CC104301C	Meeting in the Park	C01	3	3, 2, 1, 0	C1	Establishing and maintaining shared understanding
CC104305	Meeting in the Park	C01	3	1	D2	Taking appropriate action to solve the problem
CC104306	Meeting in the Park	C01	3	1	C2	Taking appropriate action to solve the problem
CC106301	Making a Film	C01	3	1	A1	Establishing and maintaining shared understanding
CC106302	Making a Film	C01	3	1	A1	Establishing and maintaining shared understanding
CC106303	Making a Film	C01	3	1	A1	Establishing and maintaining shared understanding
CC106304	Making a Film	C01	3	1	D1	Establishing and maintaining shared understanding
CC106305	Making a Film	C01	3	1	D1	Establishing and maintaining shared understanding
CC106306	Making a Film	C01	3	1	C3	Establishing and maintaining team organisation
CC106307	Making a Film	C01	3	2, 1	C2	Taking appropriate action to solve the problem
CC102101	Field Trip	C02	1	1	C3	Establishing and maintaining team organisation
CC102102C	Field Trip	C02	1	2, 1, 0	D1	Establishing and maintaining shared understanding
CC103101	Preparing a Presentation	C02	1	1	A1	Establishing and maintaining shared understanding
CC103102	Preparing a Presentation	C02	1	1	A1	Establishing and maintaining shared understanding
CC103103	Preparing a Presentation	C02	1	1	A1	Establishing and maintaining shared understanding
CC103104	Preparing a Presentation	C02	1	1	A1	Establishing and maintaining shared understanding
CC103105	Preparing a Presentation	C02	1	1	A1	Establishing and maintaining shared understanding
CC103106	Preparing a Presentation	C02	1	1	A1	Establishing and maintaining shared understanding
CC103107	Preparing a Presentation	C02	1	1	A1	Establishing and maintaining shared understanding
CC103108C	Preparing a Presentation	C02	1	4, 3, 2, 1, 0	A1	Establishing and maintaining shared understanding
CC102201	Field Trip	C02	2	1	B1	Establishing and maintaining shared understanding
CC102202	Field Trip	C02	2	1	D2	Taking appropriate action to solve the problem
CC102203	Field Trip	C02	2	1	C3	Establishing and maintaining team organisation
CC102204	Field Trip	C02	2	1	B2	Taking appropriate action to solve the problem
CC102205	Field Trip	C02	2	1	C3	Establishing and maintaining team organisation
CC102206	Field Trip	C02	2	1	B1	Establishing and maintaining shared understanding
CC102207	Field Trip	C02	2	1	C3	Establishing and maintaining team organisation
CC102209C	Field Trip	C02	2	3, 2, 1, 0	C2	Taking appropriate action to solve the problem
CC102212	Field Trip	C02	2	1	C2	Taking appropriate action to solve the problem
CC102213	Field Trip	C02	2	1	C2	Taking appropriate action to solve the problem
CC103201	Preparing a Presentation	C02	2	1	C3	Establishing and maintaining team organisation

## [Part 2/6]

**Table A.6 PISA 2015 main survey CPS item classification**

Item ID in analysis output	Unit name	C020C015 main survey cluster	Part	Score points	CPS skills	CPS competencies
CC103202	Preparing a Presentation	C02	2	1	C3	Establishing and maintaining team organisation
CC103203	Preparing a Presentation	C02	2	1	B1	Establishing and maintaining shared understanding
CC103204	Preparing a Presentation	C02	2	1	D1	Establishing and maintaining shared understanding
CC103205	Preparing a Presentation	C02	2	1	D1	Establishing and maintaining shared understanding
CC103206	Preparing a Presentation	C02	2	1	B1	Establishing and maintaining shared understanding
CC103207	Preparing a Presentation	C02	2	1	D1	Establishing and maintaining shared understanding
CC103209	Preparing a Presentation	C02	2	1	C2	Taking appropriate action to solve the problem
CC103210	Preparing a Presentation	C02	2	1	C2	Taking appropriate action to solve the problem
CC103211	Preparing a Presentation	C02	2	1	C2	Taking appropriate action to solve the problem
CC103301	Preparing a Presentation	C02	3	1	D1	Establishing and maintaining shared understanding
CC103302	Preparing a Presentation	C02	3	1	B1	Establishing and maintaining shared understanding
CC103303	Preparing a Presentation	C02	3	1	D1	Establishing and maintaining shared understanding
CC103304	Preparing a Presentation	C02	3	1	B1	Establishing and maintaining shared understanding
CC103305	Preparing a Presentation	C02	3	1	B1	Establishing and maintaining shared understanding
CC103306	Preparing a Presentation	C02	3	1	B1	Establishing and maintaining shared understanding
CC103307	Preparing a Presentation	C02	3	1	B1	Establishing and maintaining shared understanding
CC103308	Preparing a Presentation	C02	3	1	C2	Taking appropriate action to solve the problem
CC103309	Preparing a Presentation	C02	3	1	C2	Taking appropriate action to solve the problem
CC100101	Xandar	C03	1	1	C3	Establishing and maintaining team organisation
CC100102	Xandar	C03	1	1	C1	Establishing and maintaining shared understanding
CC100103	Xandar	C03	1	1	B1	Establishing and maintaining shared understanding
CC100104	Xandar	C03	1	1	B1	Establishing and maintaining shared understanding
CC100105	Xandar	C03	1	1	B3	Establishing and maintaining team organisation
CC105101	The Garden	C03	1	2, 1	B1	Establishing and maintaining shared understanding
CC105102	The Garden	C03	1	1	A1	Establishing and maintaining shared understanding
CC105103C	The Garden	C03	1	2, 1, 0	C1	Establishing and maintaining shared understanding
CC105105C	The Garden	C03	1	2, 1, 0	C2	Taking appropriate action to solve the problem
CC105108C	The Garden	C03	1	2, 1, 0	C2	Taking appropriate action to solve the problem
CC100201	Xandar	C03	2	1	A1	Establishing and maintaining shared understanding
CC100202	Xandar	C03	2	1	B3	Establishing and maintaining team organisation
CC100203	Xandar	C03	2		B3	Establishing and maintaining team organisation
CC105201C	The Garden	C03	2	2, 1, 0	A1	Establishing and maintaining shared understanding
CC105203C	The Garden	C03	2	3, 2, 1, 0	B3	Establishing and maintaining team organisation
CC105205	The Garden	C03	2	2, 1	D1	Establishing and maintaining shared understanding
CC105206	The Garden	C03	2	1	C2	Taking appropriate action to solve the problem
CC105207	The Garden	C03	2	2	D3	Establishing and maintaining team organisation
CC105208C	The Garden	C03	2	3, 2, 1, 0	C2	Taking appropriate action to solve the problem
CC105211	The Garden	C03	2	1	D3	Establishing and maintaining team organisation
CC105212C	The Garden	C03	2	2, 1, 0	C2	Taking appropriate action to solve the problem
CC105214	The Garden	C03	2	1	D3	Establishing and maintaining team organisation
CC100301	Xandar	C03	3	1	C3	Establishing and maintaining team organisation
CC100302	Xandar	C03	3	1	D1	Establishing and maintaining shared understanding
CC105301	The Garden	C03	3	1	B1	Establishing and maintaining shared understanding
CC105302	The Garden	C03	3	1	B1	Establishing and maintaining shared understanding
CC105303	The Garden	C03	3	1	D1	Establishing and maintaining shared understanding
CC105304C	The Garden	C03	3	1, 0	C1	Establishing and maintaining shared understanding
CC105306	The Garden	C03	3	1	B1	Establishing and maintaining shared understanding
CC105307	The Garden	C03	3	2, 2, 1	B1	Establishing and maintaining shared understanding
CC105308C	The Garden	C03	3	3, 2, 1, 0	D3	Establishing and maintaining team organisation
CC100401	Xandar	C03	4	1	D2	Taking appropriate action to solve the problem
CC100402	Xandar	C03	4	1	D3	Establishing and maintaining team organisation
CC105401	The Garden	C03	4	1	D1	Establishing and maintaining shared understanding
CC105402	The Garden	C03	4	1	B1	Establishing and maintaining shared understanding
CC105403	The Garden	C03	4	1	B1	Establishing and maintaining shared understanding
CC105404	The Garden	C03	4	1	B1	Establishing and maintaining shared understanding
CC105406	The Garden	C03	4	1	B3	Establishing and maintaining team organisation
CC105407	The Garden	C03	4	1	C3	Establishing and maintaining team organisation
CC105408C	The Garden	C03	4	2, 1, 0	D3	Establishing and maintaining team organisation



## [Part 3/6]

Table A.6 PISA 2015 main survey CPS item classification

Item ID in analysis output	Problem solving processes	CPS skills	Unit origin	Language of submission
CC104101	Representing and Formulating	Identifying and describing tasks to be completed	CET, Israel	English
CC104102	Exploring and Understanding	Discovering perspectives and abilities of team members	CET, Israel	English
CC104103	Exploring and Understanding	Discovering the type of collaborative interaction to solve the problem, along with goals	CET, Israel	English
CC104105	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	CET, Israel	English
CC104106	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	CET, Israel	English
CC104107	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	CET, Israel	English
CC106101	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC106102	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC106103	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC106104	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	University of Heidelberg, Germany	English
CC106105	Exploring and Understanding	Discovering the type of collaborative interaction to solve the problem, along with goals	University of Heidelberg, Germany	English
CC106106	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC106107C	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	University of Heidelberg, Germany	English
CC104201	Representing and Formulating	Describe roles and team organisation (communication protocol/rules of engagement)	CET, Israel	English
CC104202	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	CET, Israel	English
CC104203	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	CET, Israel	English
CC104204	Monitoring and reflecting	Monitoring, providing feedback and adapting the team organisation and roles	CET, Israel	English
CC104205	Representing and Formulating	Identifying and describing tasks to be completed	CET, Israel	English
CC104206	Planning and Executing	Communicating with team members about the actions to be/ being performed	CET, Israel	English
CC106201	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC106202	Representing and Formulating	Describe roles and team organisation (communication protocol/rules of engagement)	University of Heidelberg, Germany	English
CC106203	Representing and Formulating	Describe roles and team organisation (communication protocol/rules of engagement)	University of Heidelberg, Germany	English
CC106204	Representing and Formulating	Identifying and describing tasks to be completed	University of Heidelberg, Germany	English
CC106205	Representing and Formulating	Identifying and describing tasks to be completed	University of Heidelberg, Germany	English
CC106206	Monitoring and reflecting	Monitoring, providing feedback and adapting the team organisation and roles	University of Heidelberg, Germany	English
CC106207	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	University of Heidelberg, Germany	English
CC106208	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	University of Heidelberg, Germany	English
CC106209	Planning and Executing	Enacting plans	University of Heidelberg, Germany	English
CC104301C	Planning and Executing	Communicating with team members about the actions to be/ being performed	CET, Israel	English
CC104305	Monitoring and reflecting	Monitoring the results of actions and evaluating success in solving the problem	CET, Israel	English
CC104306	Planning and Executing	Enacting plans	CET, Israel	English
CC106301	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC106302	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC106303	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC106304	Monitoring and reflecting	Monitoring and repairing the shared understanding	University of Heidelberg, Germany	English
CC106305	Monitoring and reflecting	Monitoring and repairing the shared understanding	University of Heidelberg, Germany	English
CC106306	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	University of Heidelberg, Germany	English
CC106307	Planning and Executing	Enacting plans	University of Heidelberg, Germany	English
CC102101	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	GESIS, Germany	English
CC102102C	Monitoring and reflecting	Monitoring and repairing the shared understanding	GESIS, Germany	English
CC103101	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC103102	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC103103	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC103104	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC103105	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC103106	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC103107	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC103108C	Exploring and Understanding	Discovering perspectives and abilities of team members	University of Heidelberg, Germany	English
CC102201	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	GESIS, Germany	English
CC102202	Monitoring and reflecting	Monitoring the results of actions and evaluating success in solving the problem	GESIS, Germany	English
CC102203	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	GESIS, Germany	English
CC102204	Representing and Formulating	Identifying and describing tasks to be completed	GESIS, Germany	English
CC102205	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	GESIS, Germany	English
CC102206	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	GESIS, Germany	English
CC102207	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	GESIS, Germany	English
CC102209C	Planning and Executing	Enacting plans	GESIS, Germany	English
CC102212	Planning and Executing	Enacting plans	GESIS, Germany	English
CC102213	Planning and Executing	Enacting plans	GESIS, Germany	English
CC103201	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	University of Heidelberg, Germany	English

## [Part 4/6]

**Table A.6 PISA 2015 main survey CPS item classification**

Item ID in analysis output	Problem solving processes	CPS skills	Unit origin	Language of submission
CC103202	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	University of Heidelberg, Germany	English
CC103203	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	University of Heidelberg, Germany	English
CC103204	Monitoring and reflecting	Monitoring and repairing the shared understanding	University of Heidelberg, Germany	English
CC103205	Monitoring and reflecting	Monitoring and repairing the shared understanding	University of Heidelberg, Germany	English
CC103206	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	University of Heidelberg, Germany	English
CC103207	Monitoring and reflecting	Monitoring and repairing the shared understanding	University of Heidelberg, Germany	English
CC103209	Planning and Executing	Enacting plans	University of Heidelberg, Germany	English
CC103210	Planning and Executing	Enacting plans	University of Heidelberg, Germany	English
CC103211	Planning and Executing	Enacting plans	University of Heidelberg, Germany	English
CC103301	Monitoring and reflecting	Monitoring and repairing the shared understanding	University of Heidelberg, Germany	English
CC103302	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	University of Heidelberg, Germany	English
CC103303	Monitoring and reflecting	Monitoring and repairing the shared understanding	University of Heidelberg, Germany	English
CC103304	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	University of Heidelberg, Germany	English
CC103305	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	University of Heidelberg, Germany	English
CC103306	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	University of Heidelberg, Germany	English
CC103307	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	University of Heidelberg, Germany	English
CC103308	Planning and Executing	Enacting plans	University of Heidelberg, Germany	English
CC103309	Planning and Executing	Enacting plans	University of Heidelberg, Germany	English
CC100101	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	ETS, USA	English
CC100102	Planning and Executing	Communicating with team members about the actions to be/ being performed	ETS, USA	English
CC100103	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC100104	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC100105	Representing and Formulating	Describe roles and team organisation (communication protocol/rules of engagement)	ETS, USA	English
CC105101	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC105102	Exploring and Understanding	Discovering perspectives and abilities of team members	ETS, USA	English
CC105103C	Planning and Executing	Communicating with team members about the actions to be/ being performed	ETS, USA	English
CC105105C	Planning and Executing	Enacting plans	ETS, USA	English
CC105108C	Planning and Executing	Enacting plans	ETS, USA	English
CC100201	Exploring and Understanding	Discovering perspectives and abilities of team members	ETS, USA	English
CC100202	Representing and Formulating	Describe roles and team organisation (communication protocol/rules of engagement)	ETS, USA	English
CC100203	Representing and Formulating	Describe roles and team organisation (communication protocol/rules of engagement)	ETS, USA	English
CC105201C	Exploring and Understanding	Discovering perspectives and abilities of team members	ETS, USA	English
CC105203C	Representing and Formulating	Describe roles and team organisation (communication protocol/rules of engagement)	ETS, USA	English
CC105205	Monitoring and reflecting	Monitoring and repairing the shared understanding	ETS, USA	English
CC105206	Planning and Executing	Enacting plans	ETS, USA	English
CC105207	Monitoring and reflecting	Monitoring, providing feedback and adapting the team organisation and roles	ETS, USA	English
CC105208C	Planning and Executing	Enacting plans	ETS, USA	English
CC105211	Monitoring and reflecting	Monitoring, providing feedback and adapting the team organisation and roles	ETS, USA	English
CC105212C	Planning and Executing	Enacting plans	ETS, USA	English
CC105214	Monitoring and reflecting	Monitoring, providing feedback and adapting the team organisation and roles	ETS, USA	English
CC100301	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	ETS, USA	English
CC100302	Monitoring and reflecting	Monitoring and repairing the shared understanding	ETS, USA	English
CC105301	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC105302	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC105303	Monitoring and reflecting	Monitoring and repairing the shared understanding	ETS, USA	English
CC105304C	Planning and Executing	Communicating with team members about the actions to be/ being performed	ETS, USA	English
CC105306	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC105307	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC105308C	Monitoring and reflecting	Monitoring, providing feedback and adapting the team organisation and roles	ETS, USA	English
CC100401	Monitoring and reflecting	Monitoring the results of actions and evaluating success in solving the problem	ETS, USA	English
CC100402	Monitoring and reflecting	Monitoring, providing feedback and adapting the team organisation and roles	ETS, USA	English
CC105401	Monitoring and reflecting	Monitoring and repairing the shared understanding	ETS, USA	English
CC105402	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC105403	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC105404	Representing and Formulating	Building a shared representation and negotiating the meaning of the problem (common ground)	ETS, USA	English
CC105406	Representing and Formulating	Describe roles and team organisation (communication protocol/rules of engagement)	ETS, USA	English
CC105407	Planning and Executing	Following rules of engagement (e.g., prompting other team members to perform their tasks)	ETS, USA	English
CC105408C	Monitoring and reflecting	Monitoring, providing feedback and adapting the team organisation and roles	ETS, USA	English



[Part 5/6]

Table A.6 PISA 2015 main survey CPS item classification

Item ID in analysis output	International % correct	International % correct S.E.	Item parameters (RP = 0.50)						Thresholds (RP = .62)				Level
			Slope	Difficulty	Step 1	Step 2	Step 3	Step 4	1	2	3	4	
CC104101	70.03	0.21	0.926	-0.563					413				Level 1
CC104102	64.52	0.22	1.461	-0.220					458				Level 2
CC104103	56.41	0.23	1.442	-0.025					497				Level 2
CC104105	29.03	0.20	1.197	0.652					638				Level 3
CC104106	78.48	0.19	0.729	-1.010					342				Level 1
CC104107	62.06	0.21	0.824	-0.363					460				Level 2
CC106101	58.87	0.23	1.373	-0.080					488				Level 2
CC106102	30.20	0.23	0.398	1.404					882				Level 4
CC106103	55.22	0.23	1.279	-0.002					507				Level 2
CC106104	59.83	0.22	0.746	-0.216					496				Level 2
CC106105	57.59	0.22	0.933	-0.103					503				Level 2
CC106106	31.78	0.21	0.352	1.431					906				Level 4
CC106107C	50.52	0.26	0.655	0.442	0.371	-3.585	3.214		508	638	659		Level 4
CC104201	51.43	0.22	1.278	0.072					521				Level 2
CC104202	44.07	0.22	0.598	0.404					637				Level 3
CC104203	67.01	0.22	1.767	-0.225					451				Level 2
CC104204	62.43	0.22	1.504	-0.134					474				Level 2
CC104205	66.61	0.21	0.447	-0.835					425				Level 1
CC104206	78.63	0.19	1.141	-0.682					378				Level 1
CC106201	55.58	0.23	0.953	-0.014					519				Level 2
CC106202	22.32	0.19	0.356	2.270					1069				Level 4
CC106203	36.22	0.22	0.199	2.014					1144				Level 4
CC106204	35.43	0.23	1.246	0.469					600				Level 3
CC106205	56.51	0.23	1.568	-0.018					495				Level 2
CC106206	65.48	0.22	0.961	-0.321					459				Level 2
CC106207	47.36	0.23	1.198	0.168					543				Level 3
CC106208	87.65	0.16	1.561	-0.836					335				Below Level 1
CC106209	54.80	0.20	1.018	-0.005	-0.040	0.040			462	554			Level 3
CC104301C	105.30	0.23	0.954	-0.439	0.554	-0.192	-0.362		311	453	545		Level 3
CC104305	70.06	0.21	0.889	-0.518					425				Level 1
CC104306	90.78	0.13	1.791	-0.933					311				Below Level 1
CC106301	44.70	0.23	1.542	0.206					540				Level 2
CC106302	46.74	0.23	0.982	0.187					557				Level 3
CC106303	38.14	0.23	0.798	0.505					633				Level 3
CC106304	37.13	0.23	0.836	0.528					635				Level 3
CC106305	61.46	0.24	0.775	-0.276					482				Level 2
CC106306	44.83	0.24	0.223	0.706					856				Level 4
CC106307	14.41	0.15	0.799	1.060	-1.054	1.054			671	729			Level 4
CC102101	47.97	0.23	0.896	0.188					563				Level 3
CC102102C	54.27	0.19	1.003	-0.001	0.092	-0.092			463	571			Level 3
CC103101	84.35	0.17	1.254	-0.847					341				Level 1
CC103102	31.19	0.21	0.875	0.772					679				Level 4
CC103103	52.87	0.23	1.106	-0.060					502				Level 2
CC103104	64.75	0.22	1.728	-0.191					458				Level 2
CC103105	69.99	0.21	1.266	-0.420					425				Level 1
CC103106	78.68	0.21	0.497	-1.479					286				Below Level 1
CC103107	63.98	0.22	1.828	-0.182					458				Level 2
CC103108C	66.08	0.16	0.699	-0.286	-0.116	0.583	-0.562	0.096	302	406	511	574	Level 3
CC102201	32.40	0.21	0.874	0.721					670				Level 4
CC102202	34.73	0.22	0.189	2.219					1200				Level 4
CC102203	48.33	0.23	0.437	0.198					631				Level 3
CC102204	48.38	0.23	1.094	0.149					544				Level 3
CC102205	27.32	0.20	0.544	1.231					809				Level 4
CC102206	58.50	0.23	1.372	-0.064					491				Level 2
CC102207	36.47	0.23	1.796	0.389					571				Level 3
CC102209C	51.87	0.20	0.689	0.082	-0.865	0.220	0.645		424	530	567		Level 3
CC102212	28.96	0.21	0.273	2.299					1123				Level 4
CC102213	53.80	0.23	1.591	0.024					503				Level 2
CC103201	72.41	0.21	0.681	-0.723					404				Level 1

## [Part 6/6]

Table A.6 PISA 2015 main survey CPS item classification

Item ID in analysis output	International % correct	International % correct S.E.	Item parameters (RP = 0.50)					Thresholds (RP = .62)				Level	
			Slope	Difficulty	Step 1	Step 2	Step 3	Step 4	1	2	3	4	
CC103202	73.80	0.20	1.600	-0.405					418				Level 1
CC103203	49.65	0.24	1.522	0.122					524				Level 2
CC103204	63.14	0.23	1.705	-0.139					469				Level 2
CC103205	63.66	0.22	0.902	-0.266					473				Level 2
CC103206	60.77	0.23	1.878	-0.089					475				Level 2
CC103207	84.71	0.17	0.802	-1.267					284				Below Level 1
CC103209	80.04	0.18	1.382	-0.632					380				Level 1
CC103210	72.23	0.21	1.165	-0.463					420				Level 1
CC103211	69.94	0.21	1.377	-0.390					427				Level 1
CC103301	56.32	0.23	2.001	-0.011					489				Level 2
CC103302	55.70	0.23	0.384	-0.275					556				Level 3
CC103303	47.53	0.24	1.025	0.138					545				Level 3
CC103304	70.28	0.22	1.782	-0.301					435				Level 1
CC103305	64.03	0.23	1.936	-0.158					461				Level 2
CC103306	53.44	0.24	1.630	0.040					505				Level 2
CC103307	73.57	0.21	0.905	-0.620					403				Level 1
CC103308	39.64	0.23	1.456	0.341					569				Level 3
CC103309	68.99	0.22	1.740	-0.301					436				Level 1
CC100101	80.24	0.18	0.743	-1.146					314				Below Level 1
CC100102	54.45	0.22	1.350	-0.012					502				Level 2
CC100103	59.23	0.22	0.863	-0.293					471				Level 2
CC100104	50.79	0.22	1.161	0.063					524				Level 2
CC100105	68.00	0.21	1.071	-0.415					434				Level 1
CC105101	57.75	0.21	0.891	-0.118	-0.632	0.632			440	503			Level 2
CC105102	57.41	0.23	0.457	-0.328					522				Level 2
CC105103C	61.34	0.21	0.709	-0.234	-0.625	0.625			417	505			Level 2
CC105105C	49.61	0.18	1.086	0.088	0.392	-0.392			480	622			Level 3
CC105108C	8.80	0.12	0.255	3.261	-3.254	3.254			1105	1287			Level 4
CC100201	41.65	0.21	0.953	0.383					598				Level 3
CC100202	78.07	0.19	1.403	-0.622					381				Level 1
CC100203	54.78	0.22	0.327	-0.503					537				Level 2
CC105201C	49.90	0.21	0.848	0.102	-0.756	0.756			483	545			Level 3
CC105203C	39.44	0.17	0.751	0.375	-0.301	0.623	-0.322		459	593	706		Level 4
CC105205	58.24	0.12	1.097	-0.301	1.106	-1.106			565	674			Level 4
CC105206	71.67	0.22	1.197	-0.518					408				Level 1
CC105207	16.64	0.15	0.538	1.446	-0.319	0.319			747	899			Level 4
CC105208C	70.31	0.20	0.645	-0.414	-0.673	-2.441	3.113		339	440	444		Level 2
CC105211	44.26	0.23	1.167	0.203					551				Level 3
CC105212C	32.06	0.15	0.296	1.639	1.344	-1.344			785	1292			Level 4
CC105214	38.97	0.22	0.909	0.397					603				Level 3
CC100301	75.21	0.19	0.934	-0.846					357				Level 1
CC100302	16.89	0.17	0.497	2.109					992				Level 4
CC105301	44.55	0.23	1.551	0.137					526				Level 2
CC105302	74.88	0.20	0.788	-0.823					373				Level 1
CC105303	9.18	0.14	0.361	3.893					1386				Level 4
CC105304C	45.44	0.23	1.170	0.184					547				Level 3
CC105306	62.98	0.23	1.666	-0.206					456				Level 2
CC105307	31.90	0.20	0.771	0.512	-0.693	0.693			563	639			Level 3
CC105308C	34.13	0.19	0.638	0.566	-2.192	2.330	-0.138		493	620	723		Level 4
CC100401	34.33	0.23	0.527	0.814					730				Level 4
CC100402	51.42	0.23	0.475	0.058					593				Level 3
CC105401	33.85	0.24	0.513	0.997					769				Level 4
CC105402	47.73	0.24	0.729	0.143					569				Level 3
CC105403	17.93	0.19	0.322	2.847					1199				Level 4
CC105404	65.66	0.23	1.377	-0.317					442				Level 2
CC105406	87.48	0.16	0.765	-1.545					233				Below Level 1
CC105407	27.15	0.22	0.485	1.348					845				Level 4
CC105408C	39.73	0.20	0.568	0.395	-0.257	0.257			540	688			Level 4



## ANNEX B – CONTRAST CODING USED IN CONDITIONING

All tables in Annex B are available online at: [www.oecd.org/pisa](http://www.oecd.org/pisa)

## ANNEX C – STANDARD ERRORS OF MEANS, SAMPLE SIZES, SCHOOL VARIANCE ESTIMATES, AND OTHER SAMPLING OUTCOMES

**Table C.1 Standard errors of the student performance mean estimate by country and by domain**

	Science	Mathematics	Reading	Collaborative problem solving	Financial literacy
<b>OECD</b>					
Australia	1.54	1.61	1.69	1.91	1.91
Austria	2.44	2.86	2.84	2.56	
Belgium	2.29	2.35	2.42	2.39	2.88
Canada	2.08	2.31	2.30	2.27	3.75
Chile	2.38	2.54	2.58	2.69	3.17
Czech Republic	2.27	2.40	2.60	2.20	
Denmark	2.38	2.17	2.54	2.53	
Estonia	2.09	2.04	2.22	2.47	
Finland	2.39	2.31	2.55	2.55	
France	2.06	2.10	2.51	2.42	
Germany	2.69	2.89	3.01	2.85	
Greece	3.92	3.75	4.34	3.60	
Hungary	2.42	2.53	2.66	2.35	
Iceland	1.68	1.99	1.98	2.26	
Ireland	2.39	2.05	2.47		
Israel	3.44	3.63	3.78	3.62	
Italy	2.52	2.85	2.68	2.53	2.57
Japan	2.97	3.00	3.20	2.68	
Korea	3.13	3.71	3.50	2.53	
Latvia	1.56	1.87	1.80	2.26	
Luxembourg	1.12	1.27	1.44	1.50	
Mexico	2.13	2.24	2.58	2.46	
Netherlands	2.26	2.21	2.41	2.39	2.77
New Zealand	2.38	2.27	2.40	2.45	
Norway	2.26	2.23	2.51	2.52	
Poland	2.51	2.39	2.48		2.89
Portugal	2.43	2.49	2.69	2.64	
Puerto Rico (United States) <sup>1</sup>	6.09	5.55	7.11		
Slovak Republic	2.59	2.66	2.83	2.38	3.77
Slovenia	1.32	1.26	1.47	1.75	
Spain	2.07	2.15	2.36	2.15	2.88
Sweden	3.60	3.17	3.48	3.44	
Switzerland	2.90	2.92	3.03		
Turkey	3.93	4.13	3.96	3.45	
United Kingdom	2.56	2.50	2.77	2.68	
United States	3.18	3.17	3.41	3.64	3.53
<b>Partners</b>					
Albania	3.28	3.45	4.13		
Algeria	2.64	2.95	3.00		
Argentina	2.87	3.05	3.22		
Brazil	2.30	2.86	2.75	2.30	3.36
B-S-J-G (China)	4.64	4.89	5.13	3.97	5.54
Bulgaria	4.35	3.95	5.00	3.85	
Colombia	2.36	2.29	2.94	2.30	
Costa Rica	2.07	2.47	2.63	2.42	
Croatia	2.45	2.77	2.68	2.52	
Cyprus*	1.38	1.72	1.66	1.71	
Dominican Republic	2.58	2.69	3.05		
FYROM	1.25	1.28	1.41		
Georgia	2.42	2.78	2.96		
Hong Kong (China)	2.55	2.98	2.69	2.95	
Indonesia	2.57	3.08	2.87		
Jordan	2.67	2.65	2.93		
Kazakhstan	3.67	4.28	3.42		
Kosovo	1.70	1.63	1.57		
Lebanon	3.40	3.69	4.41		
Lithuania	2.65	2.33	2.74	2.46	2.97
Macao (China)	1.06	1.11	1.25	1.24	
Malaysia	3.00	3.25	3.48	3.29	
Malta	1.64	1.72	1.78		
Moldova	1.97	2.47	2.52		
Montenegro	1.03	1.46	1.58	1.27	
Peru	2.36	2.71	2.89	2.50	3.23
Qatar	1.00	1.27	1.02		
Romania	3.23	3.79	4.07		
Russia	2.91	3.11	3.08	3.42	3.19
Singapore	1.20	1.47	1.63	1.21	
Chinese Taipei	2.69	3.03	2.50	2.47	
Thailand	2.83	3.03	3.35	3.50	
Trinidad and Tobago	1.41	1.41	1.49		
Tunisia	2.10	2.95	3.06	1.94	
United Arab Emirates	2.42	2.41	2.87	2.43	
Uruguay	2.20	2.50	2.55	2.29	
Viet Nam	3.91	4.46	3.73		

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

\* Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.



[Part 1/2]

Table C.2 Sample sizes by country and by domain

	Overall student sample size	Science		Mathematics		Reading		
		School sample size	Average within-school sample size	Overall student sample size	School sample size	Reading	Overall student sample size	School sample size
<b>OECD</b>	<b>Australia</b>	14 530	758	19.17	14 530	758	19.17	14 530
	<b>Austria</b>	7 007	269	26.05	7 007	269	26.05	7 007
	<b>Belgium</b>	9 651	288	33.51	9 651	288	33.51	9 651
	<b>Canada</b>	20 058	759	26.43	20 058	759	26.43	20 058
	<b>Chile</b>	7 053	227	31.07	7 053	227	31.07	7 053
	<b>Czech Republic</b>	6 894	344	20.04	6 894	344	20.04	6 894
	<b>Denmark</b>	7 161	333	21.50	7 161	333	21.50	7 161
	<b>Estonia</b>	5 587	206	27.12	5 587	206	27.12	5 587
	<b>Finland</b>	5 882	168	35.01	5 882	168	35.01	5 882
	<b>France</b>	6 108	252	24.24	6 108	252	24.24	6 108
	<b>Germany</b>	6 522	256	25.48	6 522	256	25.48	6 522
	<b>Greece</b>	5 532	211	26.22	5 532	211	26.22	5 532
	<b>Hungary</b>	5 658	245	23.09	5 658	245	23.09	5 658
	<b>Iceland</b>	3 371	124	27.19	3 371	124	27.19	3 371
	<b>Ireland</b>	5 741	167	34.38	5 741	167	34.38	5 741
	<b>Israel</b>	6 598	173	38.14	6 598	173	38.14	6 598
	<b>Italy</b>	11 583	474	24.44	11 583	474	24.44	11 583
	<b>Japan</b>	6 647	198	33.57	6 647	198	33.57	6 647
	<b>Korea</b>	5 581	168	33.22	5 581	168	33.22	5 581
	<b>Latvia</b>	4 869	250	19.48	4 869	250	19.48	4 869
	<b>Luxembourg</b>	5 299	44	120.43	5 299	44	120.43	5 299
	<b>Mexico</b>	7 568	275	27.52	7 568	275	27.52	7 568
	<b>Netherlands</b>	5 385	187	28.80	5 385	187	28.80	5 385
	<b>New Zealand</b>	4 520	183	24.70	4 520	183	24.70	4 520
	<b>Norway</b>	5 456	229	23.83	5 456	229	23.83	5 456
	<b>Poland</b>	4 478	169	26.50	4 478	169	26.50	4 478
	<b>Portugal</b>	7 325	246	29.78	7 325	246	29.78	7 325
	<b>Puerto Rico (United States)<sup>1</sup></b>	1 398	47	29.74	1 398	47	29.74	1 398
	<b>Slovak Republic</b>	6 350	290	21.90	6 350	290	21.90	6 350
	<b>Slovenia</b>	6 406	333	19.24	6 406	333	19.24	6 406
	<b>Spain</b>	6 736	201	33.51	6 736	201	33.51	6 736
	<b>Sweden</b>	5 458	202	27.02	5 458	202	27.02	5 458
	<b>Switzerland</b>	5 860	227	25.81	5 860	227	25.81	5 860
	<b>Turkey</b>	5 895	187	31.52	5 895	187	31.52	5 895
	<b>United Kingdom</b>	14 157	550	25.74	14 157	550	25.74	14 157
	<b>United States</b>	5 712	177	32.27	5 712	177	32.27	5 712
<b>Partners</b>	<b>Albania</b>	5 215	230	22.67	5 215	230	22.67	5 215
	<b>Algeria</b>	5 519	161	34.28	5 519	161	34.28	5 519
	<b>Argentina</b>	6 349	234	27.13	6 349	234	27.13	6 349
	<b>Brazil</b>	23 141	841	27.52	23 141	841	27.52	23 141
	<b>B-S-J-G (China)</b>	9 841	268	36.72	9 841	268	36.72	9 841
	<b>Bulgaria</b>	5 928	180	32.93	5 928	180	32.93	5 928
	<b>Colombia</b>	11 795	372	31.71	11 795	372	31.71	11 795
	<b>Costa Rica</b>	6 866	205	33.49	6 866	205	33.49	6 866
	<b>Croatia</b>	5 809	160	36.31	5 809	160	36.31	5 809
	<b>Cyprus*</b>	5 571	126	44.21	5 571	126	44.21	5 571
	<b>Dominican Republic</b>	4 740	194	24.43	4 740	194	24.43	4 740
	<b>FYROM</b>	5 324	106	50.23	5 324	106	50.23	5 324
	<b>Georgia</b>	5 316	262	20.29	5 316	262	20.29	5 316
	<b>Hong Kong (China)</b>	5 359	138	38.83	5 359	138	38.83	5 359
	<b>Indonesia</b>	6 513	236	27.60	6 513	236	27.60	6 513
	<b>Jordan</b>	7 267	250	29.07	7 267	250	29.07	7 267
	<b>Kazakhstan</b>	4 826	224	21.54	4 826	224	21.54	4 826
	<b>Kosovo</b>	7 841	232	33.80	7 841	232	33.80	7 841
	<b>Lebanon</b>	4 546	270	16.84	4 546	270	16.84	4 546
	<b>Lithuania</b>	6 525	311	20.98	6 525	311	20.98	6 525
	<b>Macao (China)</b>	4 476	45	99.47	4 476	45	99.47	4 476
	<b>Malaysia</b>	8 861	225	39.38	8 861	225	39.38	8 861
	<b>Malta</b>	3 634	59	61.59	3 634	59	61.59	3 634
	<b>Moldova</b>	5 325	229	23.25	5 325	229	23.25	5 325
	<b>Montenegro</b>	5 665	64	88.52	5 665	64	88.52	5 665
	<b>Peru</b>	6 971	281	24.81	6 971	281	24.81	6 971
	<b>Qatar</b>	12 083	167	72.35	12 083	167	72.35	12 083
	<b>Romania</b>	4 876	182	26.79	4 876	182	26.79	4 876
	<b>Russia</b>	6 036	210	28.74	6 036	210	28.74	6 036
	<b>Singapore</b>	6 115	177	34.55	6 115	177	34.55	6 115
	<b>Chinese Taipei</b>	7 708	214	36.02	7 708	214	36.02	7 708
	<b>Thailand</b>	8 249	273	30.22	8 249	273	30.22	8 249
	<b>Trinidad and Tobago</b>	4 692	149	31.49	4 692	149	31.49	4 692
	<b>Tunisia</b>	5 375	165	32.58	5 375	165	32.58	5 375
	<b>United Arab Emirates</b>	14 167	473	29.95	14 167	473	29.95	14 167
	<b>Uruguay</b>	6 062	220	27.55	6 062	220	27.55	6 062
	<b>Viet Nam</b>	5 826	188	30.99	5 826	188	30.99	5 826

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

\* See note 1 under Table C.1.



[Part 2/2]

Table C.2 Sample sizes by country and by domain

	Collaborative problem solving			Financial literacy		
	Overall student sample size	School sample size	Average within-school sample size	Overall student sample size	School sample size	Average within-school sample size
<b>OECD</b>						
Australia	14 530	758	19.17	14 530	758	19.17
Austria	7 007	269	26.05			
Belgium	9 651	288	33.51	5 675	175	32.43
Canada	20 058	759	26.43	13 082	487	26.86
Chile	7 053	227	31.07	7 053	227	31.07
Czech Republic	6 894	344	20.04			
Denmark	7 161	333	21.50			
Estonia	5 587	206	27.12			
Finland	5 882	168	35.01			
France	6 108	252	24.24			
Germany	6 522	256	25.48			
Greece	5 532	211	26.22			
Hungary	5 658	245	23.09			
Iceland	3 371	124	27.19			
Ireland						
Israel	5 988	154	38.88			
Italy	11 583	474	24.44	11 583	474	24.44
Japan	6 647	198	33.57			
Korea	5 581	168	33.22			
Latvia	4 869	250	19.48			
Luxembourg	5 299	44	120.43			
Mexico	7 568	275	27.52			
Netherlands	5 385	187	28.80	5 385	187	28.80
New Zealand	4 520	183	24.70			
Norway	5 456	229	23.83			
Poland				4 478	169	26.50
Portugal	7 325	246	29.78			
Puerto Rico (United States) <sup>1</sup>						
Slovak Republic	6 350	290	21.90	6 350	290	21.90
Slovenia	6 406	333	19.24			
Spain	6 736	201	33.51	6 736	201	33.51
Sweden	5 458	202	27.02			
Switzerland						
Turkey	5 895	187	31.52			
United Kingdom	14 157	550	25.74			
United States	5 712	177	32.27	5 712	177	32.27
<b>Partners</b>						
Albania						
Algeria						
Argentina						
Brazil	23 141	841	27.52	23 141	841	27.52
B-S-J-G (China)	9 841	268	36.72	9 841	268	36.72
Bulgaria	5 928	180	32.93			
Colombia	11 795	372	31.71			
Costa Rica	6 866	205	33.49			
Croatia	5 809	160	36.31			
Cyprus*	5 571	126	44.21			
Dominican Republic						
FYROM						
Georgia						
Hong Kong (China)	5 359	138	38.83			
Indonesia						
Jordan						
Kazakhstan						
Kosovo						
Lebanon						
Lithuania	6 525	311	20.98	6 525	311	20.98
Macao (China)	4 476	45	99.47			
Malaysia	8 861	225	39.38			
Malta						
Moldova						
Montenegro	5 665	64	88.52			
Peru	6 971	281	24.81	6 971	281	24.81
Qatar						
Romania						
Russia	6 036	210	28.74	6 036	210	28.74
Singapore	6 115	177	34.55			
Chinese Taipei	7 708	214	36.02			
Thailand	8 249	273	30.22			
Trinidad and Tobago						
Tunisia	5 375	165	32.58			
United Arab Emirates	14 167	473	29.95			
Uruguay	6 062	220	27.55			
Viet Nam						

1. Puerto Rico is an unincorporated territory of the United States. As such, PISA results for the United States do not include Puerto Rico.

\* See note 1 under Table C.1.



Table C.3 School variance estimate by country and by domain

	Reading	Mathematics	Science
	Variance	Variance	Variance
<b>OECD</b>			
Australia	10 544.7	8 660.5	10 465.0
Austria	10 229.4	9 052.2	9 475.9
Belgium	10 043.9	9 478.8	10 037.2
Canada	8 610.8	7 698.4	8 531.7
Chile	7 770.3	7 300.1	7 399.2
Czech Republic	10 091.8	8 224.3	9 075.4
Denmark	7 621.4	6 491.9	8 153.0
Estonia	7 655.2	6 467.4	7 903.6
Finland	8 812.8	6 749.4	9 249.6
France	12 552.5	9 056.1	10 396.6
Germany	10 024.2	7 919.8	9 866.4
Greece	9 643.1	7 987.9	8 450.1
Hungary	9 415.6	8 798.4	9 280.9
Iceland	9 874.2	8 634.1	8 318.8
Ireland	7 428.5	6 364.4	7 902.9
Israel	12 794.3	10 684.4	11 313.5
Italy	8 796.5	8 755.2	8 361.3
Japan	8 545.3	7 784.3	8 737.1
Korea	9 416.1	9 945.0	9 059.2
Latvia	7 188.5	6 012.3	6 758.5
Luxembourg	11 371.4	8 752.8	10 080.6
Mexico	6 088.9	5 626.6	5 098.9
Netherlands	10 198.8	8 376.8	10 189.3
New Zealand	11 035.1	8 485.6	10 836.0
Norway	9 759.7	7 208.7	9 262.6
Poland	8 025.2	7 681.5	8 243.6
Portugal	8 455.1	9 167.8	8 431.4
Slovak Republic	10 865.0	9 106.1	9 787.5
Slovenia	8 420.4	7 715.9	9 061.2
Spain	7 629.6	7 177.1	7 745.7
Sweden	10 360.0	8 114.1	10 502.1
Switzerland	9 579.8	9 168.1	9 905.3
Turkey	6 789.7	6 711.1	6 282.6
United Kingdom	9 349.1	8 568.6	9 930.5
United States	9 970.4	7 826.5	9 726.5
<b>OECD average</b>	9 284.5	8 050.1	8 966.3
<b>Partners</b>			
Albania	9 337.2	7 436.8	6 159.1
Algeria	5 284.9	5 060.5	4 800.1
Brazil	10 038.1	7 952.6	7 948.0
B-S-J-G (China)	11 864.2	11 246.5	10 688.8
Bulgaria	13 134.3	9 445.6	10 307.3
CABA (Argentina)	8 190.6	7 836.0	7 356.3
Colombia	8 072.9	5 956.4	6 460.0
Costa Rica	6 279.6	4 683.8	4 903.3
Croatia	8 234.8	7 796.4	7 977.9
Cyprus*	10 464.7	8 540.9	8 617.6
Dominican Republic	7 207.0	4 697.2	5 251.7
FYROM	9 839.8	9 204.6	7 188.0
Georgia	10 742.4	8 816.7	8 207.8
Hong Kong (China)	7 365.2	8 126.4	6 492.1
Indonesia	5 780.1	6 372.5	4 674.6
Jordan	8 853.5	7 369.4	7 121.2
Kosovo	6 132.7	5 675.6	5 082.1
Lebanon	13 331.8	10 227.5	8 173.6
Lithuania	8 911.2	7 482.3	8 266.6
Macao (China)	6 744.6	6 385.4	6 622.0
Malta	14 533.3	12 162.2	13 839.4
Moldova	9 572.8	8 126.9	7 402.7
Montenegro	8 850.5	7 506.8	7 268.5
Peru	7 946.5	6 822.8	5 882.7
Qatar	12 240.1	9 760.4	9 749.1
Romania	9 037.7	7 444.6	6 258.7
Russia	7 641.6	6 907.6	6 792.2
Singapore	9 745.6	9 104.7	10 733.7
Chinese Taipei	8 687.2	10 596.2	9 911.2
Thailand	6 372.7	6 643.0	6 160.1
Trinidad and Tobago	10 831.1	9 212.8	8 797.9
Tunisia	6 657.7	7 092.5	4 206.1
United Arab Emirates	11 179.1	9 318.9	9 828.4
Uruguay	9 330.2	7 498.0	7 490.1
Viet Nam	5 271.0	7 011.4	5 867.6
Argentina	7 890.9	6 497.4	6 496.4
Kazakhstan	6 473.7	6 790.3	5 841.5
Malaysia	6 555.6	6 415.8	5 734.8

\* See note 1 under Table C.1.

Table C.4 Intraclass correlation by country and by domain

	Reading	Mathematics	Science
	Rho	Rho	Rho
<b>OECD</b>			
Australia	0.22	0.22	0.22
Austria	0.47	0.45	0.44
Belgium	0.46	0.46	0.45
Canada	0.18	0.21	0.16
Chile	0.38	0.39	0.39
Czech Republic	0.45	0.46	0.45
Denmark	0.16	0.13	0.14
Estonia	0.19	0.18	0.19
Finland	0.10	0.09	0.08
France	0.52	0.50	0.50
Germany	0.46	0.43	0.45
Greece	0.39	0.32	0.36
Hungary	0.57	0.53	0.55
Iceland	0.05	0.05	0.04
Ireland	0.13	0.15	0.13
Israel	0.43	0.41	0.37
Italy	0.43	0.41	0.43
Japan	0.43	0.47	0.44
Korea	0.27	0.27	0.25
Latvia	0.18	0.16	0.17
Luxembourg	0.31	0.31	0.34
Mexico	0.34	0.29	0.30
Netherlands	0.58	0.58	0.58
New Zealand	0.18	0.19	0.18
Norway	0.10	0.09	0.08
Poland	0.14	0.13	0.14
Portugal	0.25	0.24	0.23
Slovak Republic	0.49	0.42	0.45
Slovenia	0.50	0.46	0.48
Spain	0.15	0.14	0.13
Sweden	0.16	0.17	0.16
Switzerland	0.37	0.34	0.38
Turkey	0.52	0.50	0.53
United Kingdom	0.21	0.23	0.22
United States	0.19	0.21	0.19
<b>OECD average</b>	0.31	0.30	0.30
<b>Partners</b>			
Albania	0.25	0.24	0.24
Algeria	0.33	0.31	0.31
Brazil	0.38	0.42	0.39
B-S-J-G (China)	0.54	0.52	0.53
Bulgaria	0.52	0.47	0.52
CABA (Argentina)	0.34	0.43	0.35
Colombia	0.35	0.30	0.33
Costa Rica	0.32	0.29	0.29
Croatia	0.41	0.37	0.37
Cyprus*	0.26	0.27	0.24
Dominican Republic	0.41	0.37	0.37
FYROM	0.31	0.29	0.28
Georgia	0.26	0.28	0.23
Hong Kong (China)	0.32	0.31	0.31
Indonesia	0.40	0.45	0.42
Jordan	0.34	0.27	0.27
Kosovo	0.34	0.30	0.30
Lebanon	0.52	0.48	0.48
Lithuania	0.36	0.29	0.34
Macao (China)	0.26	0.19	0.23
Malta	0.33	0.30	0.30
Moldova	0.22	0.21	0.20
Montenegro	0.29	0.26	0.26
Peru	0.46	0.36	0.36
Qatar	0.45	0.42	0.40
Romania	0.41	0.40	0.39
Russia	0.20	0.20	0.19
Singapore	0.35	0.34	0.35
Chinese Taipei	0.33	0.36	0.36
Thailand	0.37	0.31	0.34
Trinidad and Tobago	0.53	0.57	0.54
Tunisia	0.40	0.35	0.37
United Arab Emirates	0.46	0.43	0.42
Uruguay	0.38	0.36	0.36
Viet Nam	0.48	0.43	0.40
Argentina	0.32	0.33	0.30
Kazakhstan	0.39	0.48	0.46
Malaysia	0.30	0.30	0.27

\* See note 1 under Table C.1.

Note: The intraclass correlation measures the variation in student performance accounted for by clustering, i.e. is the ratio of the between-school variance and the sum of the between-school and within-school variance. For further details on how multilevel models are calibrated in PISA, please refer to Annex 3 of OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris.



Table C.5 Within explicit strata intraclass correlation by country and by domain

	Reading		Mathematics		Science	
	Rho	S.E.	Rho	S.E.	Rho	S.E.
<b>OECD</b>						
Australia	0.20	(0.02)	0.20	(0.02)	0.19	(0.02)
Austria	0.12	(0.02)	0.18	(0.02)	0.14	(0.02)
Belgium	0.21	(0.02)	0.22	(0.02)	0.22	(0.02)
Canada	0.17	(0.01)	0.18	(0.01)	0.14	(0.00)
Chile	0.22	(0.02)	0.23	(0.02)	0.22	(0.02)
Czech Republic	0.14	(0.02)	0.14	(0.02)	0.16	(0.02)
Denmark	0.13	(0.01)	0.10	(0.01)	0.11	(0.01)
Estonia	0.17	(0.02)	0.15	(0.02)	0.15	(0.02)
Finland	0.10	(0.02)	0.10	(0.02)	0.08	(0.02)
France	0.19	(0.03)	0.21	(0.03)	0.19	(0.03)
Germany	0.42	(0.03)	0.42	(0.03)	0.42	(0.02)
Greece	0.40	(0.03)	0.36	(0.03)	0.39	(0.03)
Hungary	0.31	(0.03)	0.30	(0.03)	0.33	(0.03)
Iceland	0.04	(0.01)	0.04	(0.01)	0.03	(0.01)
Ireland	0.09	(0.01)	0.11	(0.02)	0.09	(0.01)
Israel	0.25	(0.04)	0.24	(0.03)	0.24	(0.04)
Italy	0.23	(0.01)	0.29	(0.01)	0.29	(0.01)
Japan	0.44	(0.03)	0.47	(0.03)	0.44	(0.03)
Korea	0.19	(0.02)	0.19	(0.02)	0.17	(0.02)
Latvia	0.14	(0.02)	0.14	(0.02)	0.14	(0.02)
Luxembourg	0.24	(0.04)	0.24	(0.04)	0.24	(0.04)
Mexico	0.27	(0.03)	0.23	(0.02)	0.25	(0.03)
Netherlands	0.28	(0.04)	0.29	(0.04)	0.24	(0.02)
New Zealand	0.19	(0.02)	0.18	(0.02)	0.18	(0.02)
Norway	0.10	(0.01)	0.09	(0.01)	0.08	(0.01)
Poland	0.14	(0.02)	0.13	(0.02)	0.15	(0.02)
Portugal	0.18	(0.02)	0.19	(0.02)	0.17	(0.01)
Slovak Republic	0.31	(0.03)	0.27	(0.03)	0.27	(0.02)
Slovenia	0.17	(0.01)	0.20	(0.02)	0.20	(0.02)
Spain	0.09	(0.01)	0.08	(0.02)	0.08	(0.02)
Sweden	0.14	(0.02)	0.15	(0.02)	0.14	(0.02)
Switzerland	0.19	(0.01)	0.19	(0.01)	0.20	(0.02)
Turkey	0.33	(0.03)	0.36	(0.03)	0.35	(0.03)
United Kingdom	0.15	(0.01)	0.15	(0.01)	0.15	(0.00)
United States	0.19	(0.02)	0.20	(0.02)	0.18	(0.02)
<b>OECD average</b>	<b>0.20</b>	<b>(0.00)</b>	<b>0.21</b>	<b>(0.00)</b>	<b>0.20</b>	<b>(0.00)</b>
<b>Partners</b>						
Albania	0.19	(0.02)	0.21	(0.02)	0.20	(0.02)
Algeria	0.35	(0.03)	0.30	(0.04)	0.34	(0.03)
Brazil	0.31	(0.03)	0.37	(0.04)	0.34	(0.03)
B-S-J-G (China)	0.31	(0.01)	0.29	(0.01)	0.29	(0.01)
Bulgaria	0.52	(0.03)	0.47	(0.03)	0.51	(0.03)
CABA (Argentina)	0.38	(0.05)	0.44	(0.04)	0.38	(0.04)
Colombia	0.25	(0.03)	0.19	(0.02)	0.22	(0.02)
Costa Rica	0.18	(0.02)	0.19	(0.02)	0.15	(0.02)
Croatia	0.21	(0.02)	0.19	(0.02)	0.19	(0.02)
Cyprus*	0.24	(0.03)	0.23	(0.03)	0.19	(0.02)
Dominican Republic	0.20	(0.03)	0.18	(0.03)	0.18	(0.03)
FYROM	0.25	(0.03)	0.24	(0.03)	0.21	(0.03)
Georgia	0.18	(0.03)	0.22	(0.03)	0.17	(0.02)
Hong Kong (China)	0.33	(0.03)	0.31	(0.03)	0.32	(0.03)
Indonesia	0.34	(0.03)	0.38	(0.03)	0.36	(0.03)
Jordan	0.32	(0.03)	0.23	(0.03)	0.25	(0.03)
Kosovo	0.27	(0.03)	0.27	(0.03)	0.23	(0.03)
Lebanon	0.40	(0.03)	0.35	(0.04)	0.34	(0.03)
Lithuania	0.18	(0.02)	0.15	(0.02)	0.18	(0.02)
Macao (China)	0.27	(0.03)	0.25	(0.04)	0.28	(0.03)
Malta	0.13	(0.02)	0.12	(0.02)	0.12	(0.02)
Moldova	0.08	(0.01)	0.12	(0.02)	0.09	(0.01)
Montenegro	0.23	(0.05)	0.22	(0.05)	0.20	(0.04)
Peru	0.32	(0.02)	0.23	(0.02)	0.23	(0.02)
Qatar	0.30	(0.01)	0.24	(0.02)	0.25	(0.02)
Romania	0.41	(0.03)	0.41	(0.03)	0.40	(0.03)
Russia	0.13	(0.02)	0.14	(0.02)	0.13	(0.02)
Singapore	0.31	(0.03)	0.31	(0.03)	0.32	(0.03)
Chinese Taipei	0.15	(0.02)	0.15	(0.02)	0.15	(0.02)
Thailand	0.29	(0.03)	0.22	(0.02)	0.24	(0.03)
Trinidad and Tobago	0.43	(0.03)	0.45	(0.02)	0.42	(0.02)
Tunisia	0.37	(0.04)	0.33	(0.04)	0.35	(0.04)
United Arab Emirates	0.33	(0.02)	0.24	(0.02)	0.25	(0.02)
Uruguay	0.14	(0.02)	0.15	(0.02)	0.13	(0.01)
Viet Nam	0.45	(0.03)	0.40	(0.03)	0.37	(0.03)
Argentina	0.30	(0.02)	0.29	(0.03)	0.28	(0.02)
Kazakhstan	0.26	(0.03)	0.39	(0.04)	0.33	(0.04)
Malaysia	0.23	(0.03)	0.23	(0.03)	0.19	(0.02)

\* See note 1 under Table C.1.

Note: The within explicit strata intraclass correlation has been computed from a multilevel model where Level 2 (i.e. school) weights correspond to the sum of final student weights ( $W_{FSTUWT}$ ) within each stratum. For further details on how multilevel models are calibrated in PISA, please refer to Annex 3 of OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris.



## [Part 1/1]

Table C.6 Percentage of school variance explained by explicit stratification variables by country and by domain

	Reading	Mathematics	Science
	% Variance	% Variance	% Variance
<b>OECD</b>			
Australia	15.84	16.54	16.14
Austria	84.77	74.96	80.37
Belgium	62.27	62.83	62.10
Canada	5.51	15.15	8.20
Chile	49.71	54.09	53.08
Czech Republic	79.80	78.20	73.84
Denmark	17.29	22.45	23.60
Estonia	6.01	6.19	11.16
Finland	10.63	11.24	12.26
France	53.49	47.32	50.63
Germany	23.95	19.00	19.30
Greece	1.58	0.90	1.20
Hungary	58.96	55.18	53.28
Iceland	9.43	10.73	9.66
Ireland	16.81	14.09	15.95
Israel	56.30	55.53	42.16
Italy	59.60	43.95	46.14
Japan	3.61	5.13	4.27
Korea	25.03	26.50	26.10
Latvia	22.84	14.93	18.50
Luxembourg	19.19	19.33	22.14
Mexico	22.73	23.09	20.32
Netherlands	71.97	71.46	76.89
New Zealand	5.39	8.77	6.90
Norway	0.00	0.01	0.01
Poland	0.00	0.01	0.00
Portugal	24.53	19.02	24.21
Slovak Republic	40.99	37.03	41.63
Slovenia	75.10	64.41	69.88
Spain	41.80	47.31	45.71
Sweden	9.34	8.22	9.37
Switzerland	36.93	33.05	38.43
Turkey	46.80	40.24	46.47
United Kingdom	13.88	17.10	19.43
United States	4.76	5.11	5.74
<b>OECD average</b>	<b>30.77</b>	<b>29.40</b>	<b>30.14</b>
<b>Partners</b>			
Albania	21.59	12.37	15.53
Algeria	5.19	7.46	6.07
Brazil	13.42	11.50	12.71
B-S-J-G (China)	52.35	51.82	54.68
Bulgaria	13.44	14.87	15.42
CABA (Argentina)	0.00	0.00	0.00
Colombia	12.08	12.61	13.01
Costa Rica	32.93	27.43	35.79
Croatia	53.79	53.54	52.19
Cyprus*	30.85	37.29	44.05
Dominican Republic	55.62	51.00	47.11
FYROM	11.01	16.12	23.93
Georgia	32.82	25.69	29.59
Hong Kong (China)	0.08	0.18	0.14
Indonesia	7.84	9.63	8.98
Jordan	6.29	9.43	8.26
Kosovo	7.25	5.78	7.08
Lebanon	42.81	45.24	45.66
Lithuania	47.27	40.61	41.88
Macao (China)	3.56	2.14	2.92
Malta	63.65	64.88	65.56
Moldova	47.91	30.33	39.84
Montenegro	23.94	19.17	24.55
Peru	25.19	24.84	27.65
Qatar	50.98	60.52	53.86
Romania	1.29	0.51	0.65
Russia	42.20	31.93	40.08
Singapore	14.67	11.88	12.79
Chinese Taipei	59.36	63.90	63.27
Thailand	52.42	60.05	60.90
Trinidad and Tobago	42.19	42.77	43.04
Tunisia	13.41	15.94	13.80
United Arab Emirates	45.34	58.94	56.04
Uruguay	67.02	64.87	67.06
Viet Nam	13.68	9.00	9.96
Argentina	9.29	8.72	8.23
Kazakhstan	45.82	29.42	39.34
Malaysia	6.28	6.25	8.58

\* See note 1 under Table C.1.

Note: The percentage of school variance explained by explicit stratification variables has been computed from a multilevel model where Level 2 (i.e. school) weights correspond to the sum of final student weights ( $W_{FSTUWT}$ ) within each stratum. This percentage corresponds to the ratio of 1. the difference of the between-school variance from an "empty" model and the between-school variance from the "full" model with dummy variables for the STRATUM variable and 2. the between-school variance from this "full" model. It thus represents the reduction in between-school variation accounted for by explicit stratification variables. For further details on how multilevel models are calibrated in PISA, please refer to Annex 3 of OECD (2016), *PISA 2015 Results (Volume I): Excellence and Equity in Education*, PISA, OECD Publishing, Paris.



## ANNEX D – MAPPING OF ISCED TO YEARS

Table D.1 Mapping of ISCED to years

	Completed ISCED Level 0 (pre-primary education)	Completed ISCED Level 1 (primary education)	Completed ISCED Level 2 (lower secondary education)	Completed ISCED Levels 3B or 3C (upper secondary education providing direct access to the labour market or to ISCED 5B programmes)	Completed ISCED Level 3A (upper secondary education providing access to ISCED 5A and 5B programmes) and/or ISCED Level 4 (non-tertiary post-secondary)	Completed ISCED Level 5B (non-university tertiary education)	Completed ISCED Level 5A (university level tertiary education) or ISCED Level 6 (advanced research programmes)
<b>OECD</b>							
Australia	3	6	10	11	12	14	15
Austria	3	4	9	12	12.5	15	17
Belgium	3	6	9	12	12	15	17
Canada	3	6	9	12	12	15	17
Chile	3	6	8	12	12	16	17
Czech Republic	3	5	9	11	13	16	16
Denmark	3	7	10	13	13	16	18
Estonia	3	6	9	12	12	15	16
Finland	3	6	9	12	12	14.5	16.5
France	3	5	9	12	12	14	15
Germany	3	4	10	13	13	15	18
Greece	3	6	9	11.5	12	15	17
Hungary	3	4	8	10.5	12	13.5	16.5
Iceland	3	7	10	13	14	16	18
Ireland	3	6	9	12	12	14	16
Israel	3	6	9	12	12	15	15
Italy	3	5	8	12	13	16	17
Japan	3	6	9	12	12	14	16
Korea	3	6	9	12	12	14	16
Latvia	3	4	8	11	11	14	16
Luxembourg	3	6	9	12	13	16	17
Mexico	3	6	9	12	12	14	16
Netherlands	3	6	10	13	12	15	16
New Zealand	3	5.5	10	11	12	14	15
Norway	3	6	9	12	12	14	16
Poland	3	.	8	11	12	15	16
Portugal	3	6	9	12	12	15	17
Slovak Republic	3	4	9	12	13	16	18
Slovenia	3	4	8	11	12	15	16
Spain	3	5	8	10	12	13	16.5
Sweden	3	6	9	11.5	12	14	16
Switzerland	3	6	9	12.5	12.5	14.5	17.5
Turkey	3	4	8	12	12	14	16
United Kingdom	3	7	10	11	12	14	16
United States	3	6	9	12	12	14	16
<b>Partners</b>							
Albania	3	6	10	12	12	16	16
Algeria	3	5	9	11	12	12	15
Argentina	3	6	10	12	12	14.5	17
Brazil	3	5	9	12	12	14.5	17
B-S-J-G (China)	3	6	9	12	12	15	16.5
Bulgaria	3	4	8	10	12	15	17.5
Colombia	3	5	9	11	11	14	15.5
Costa Rica	3	6	9	11	12	14	16
Croatia	3	4	8	11	12	15	17
Cyprus*	3	6	9	12	12	15	16.5
Dominican Republic	3	6	9	11	12	14	16
FYROM	3	5	9	13	13	15	17
Georgia	3	6	9	11	12	13	15.5
Hong Kong (China)	3	6	9	11	13	14	16
Indonesia	3	6	9	12	12	14	15
Jordan	3	6	10	12	12	14.5	16
Kazakhstan	3	4	9	11.5	12.5	14	15
Kosovo	3	5	9	12	14	16	18
Lebanon	3	6	9	12	12	15	16
Lithuania	3	3	8	11	11	15	16
Macao (China)	3	6	9	11	12	15	16
Malaysia	3	6	9	11	13	16	18
Malta	3	6	9	12	13	15	17
Moldova	3	4	9	11	12	14	16.5
Montenegro	3	4	8	11	12	15	16
Peru	3	6	9	11	11	14	17
Qatar	3	6	9	12	12	15	16
Romania	3	4	8	11.5	12.5	14	16
Russia	3	4	9	11	11	13	16
Singapore	3	6	8	10	11	13	16
Chinese Taipei	3	6	9	12	12	14	16
Thailand	3	6	9	12	12	14	16
Trinidad and Tobago	3	5	9	12	12	15	16
Tunisia	3	6	9	12	13	16	17
United Arab Emirates	3	5	9	12	12	15	16
Uruguay	3	6	9	12	12	15	17
Viet Nam	3	5	9	12	12	15	17

\* Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.



## ANNEX E – NATIONAL HOUSEHOLD POSSESSION ITEMS

[Part 1/2]

Table E.1 National household possession items

	ST011Q17TA	ST011Q18TA	ST011Q19TA
<b>OECD</b>			
Australia	Solar panels (on the roof)	A home gym and/or gym membership	Espresso machine
Austria	A laptop/notebook of your own	Electronic devices for playing (Playstation®, Nintendo®, X-Box®, Wii®)	A swimming pool
Belgium	A home cinema set (LCD or LED screen with home cinema system)	An alarm system	A housekeeper
Canada	iPod®/An MP3 player	A subscription to a daily newspaper	Air conditioning
Chile	A second house (vacation house)	A digital video camera	Microwave oven
Czech Republic	N/A	N/A	N/A
Denmark	A musical instrument (e.g. piano, guitar, violin)	A smart TV	N/A
Estonia	Video camera	Digital camera	Dishwasher
Finland	A laptop	Home alarm system	N/A
France	An espresso machine	A digital camera (not installed in a mobile phone)	A Hi-Fi system
Germany	electronical devices for playing (Playstation®, Nintendo®, X-Box®, Wii®)	A TV in your own room	Audiobooks
Greece	Dishwasher	Garage or parking space	Alarm system
Hungary	Video games console (e.g. Playstation®)	Tablet computer (e.g. iPad®, Samsung®)	Digital camera (not part of a phone)
Iceland	Security watch or system	Hot tub	House cleaning service.
Ireland	Your own MP3 player (e.g. iPod®)	Your own laptop or tablet computer (e.g. iPad®, BlackBerry®, PlayBook™)	Your own smartphone (e.g. iPhone®, Samsung® or Sony® android phone)
Israel	4x4 vehicle	Espresso machine	Home cinema system
Italy	Antique furniture	Alarm system	Air-conditioning
Japan	Digital camera	Smart Phones	Clothing Dryer
Korea	Air conditioner	Home theatre	Dishwasher
Latvia	Personal smartphone	Bicycle	Scooter
Luxembourg	A smartphone with unlimited Internet access	iPad®	A recent game console (e.g. Playstation 4® or Wii U®)
Mexico	A BluRay player	Phone line	Microwave oven
Netherlands	An alarm system on the house	A piano	Energy regulator
New Zealand	Heat pump	A boat (e.g. a leisure craft, yacht)	Snow skis
Norway	iPad®	iPhone®	N/A
Poland	Dishwasher	Digital camera	Printer
Portugal	Home cinema	Central heating	Plasma or LCD TV
Slovak Republic	Video camera	Digital camera (not as a part of a mobile phone, but separate one)	N/A
Slovenia	Your own computer	Attending an extra out-of-school-time activities paid by your parents	Travelling abroad for one week or more
Spain	Video camera	A tablet (iPad®, Samsung®)	Home cinema
Sweden	Piano	Jacuzzi	Espresso machine
Switzerland	Musical instrument (excluding Recorder)	An iPhone®	A digital video camera
Turkey	Air conditioning type heating-cooling system	Video camera	Home Theatre system
United Kingdom (excl. Scotland)	A premium TV package (e.g. Sky movies, Sky sports)	A high definition (HD) TV	A tablet computer (e.g. iPad®)
United Kingdom (Scotland)	A premium TV package (e.g. Sky movies, Sky sports)	A musical instrument (e.g. piano, violin)	Do your parents pay for a cleaner to clean your home?
United States	A guest room	A high-speed Internet connection	A musical instrument



[Part 2/2]

Table E.1 National household possession items

	ST011Q17TA	ST011Q18TA	ST011Q19TA
<b>Partners</b>			
Albania	Microwave	Cultural television programs with payment	Digital camera
Algeria	N/A	N/A	N/A
Argentina	Dishwasher	Air conditioning	Washing machine
Brazil	Blue-Ray Player	Video game	iPod®
B-S-J-G (China)	Vacuum cleaner	Digital camera or point-and-shoot	Juice maker
Bulgaria	Smart phone	Digital camera	Air conditioning
Colombia	Digital camera	N/A	Encyclopaedia
Costa Rica	Cable TV	A console of video games	A home theatre set
Croatia	Laundry dryer	Game console (e.g. Playstation 3® or Nintendo Wii®)	Air conditioner
Cyprus*	A Home Cinema	A Jacuzzi	Home security alarm system
Dominican Republic	Microwave	Air conditioning	Decoration objects
FYROM	LCD projector	Interactive whiteboard	N/A
Georgia	Dishwasher	Family cinema	Swimming pool
Hong Kong (China)	Violin / Cello	Piano	Golf equipment
Indonesia	Digital camera	Motorcycle	Car
Jordan	Central heating	Playstation®	Digital camera
Kazakhstan	Digital fotocamera	Video camera	Satellite antenna
Kosovo	Central heating	Cultural television programs	Digital camera
Lebanon	DVD player	Flat screen TV \ plasma \ LCD	Cable TV \ paid \ satellite
Lithuania	Digital camera	Press, Subscription edition (newspaper, magazine)	Cinecamera
Macao (China)	A piano or violin	A digital camera	An iPad®
Malaysia	Television	Refrigerator	Air conditioner
Malta	Photovoltaic panels	Summer residence	Swimming pool
Moldova	Laptop	N/A	N/A
Montenegro	Cable TV	Plasma TV	Digital camera
Peru	Video games (Play Station®, Nintendo®, Wii®)	Refrigerator	Washing machine
Qatar	MP3 Walkman	Digital video camera	Video games console
Romania	Plasma or LCD TV	Cable / satellite TV	Digital video camera
Russia	Jacuzzi	Home cinema	House or cottage constructed during the last 15 years
Singapore	Air-conditioner	Domestic helper (e.g. full/part-time maid)	N/A
Chinese Taipei	Piano, violin	iPod®	Digital camera
Thailand	Air conditioning	Electric massage chair	Microwave Oven
Trinidad and Tobago	Flat screen TV/Plasma TV/LCD TV/ Smart TV	A DVD player	Refrigerator with ice maker
Tunisia	Flat screen TV	Digital camera	Washing machine
United Arab Emirates	A laptop of your own	Electronic games (Wii®, Xbox®)	iPad®
Uruguay	A dishwasher	Refrigerator with freezer	Notebook PC or laptop (XO Ceibal not included)
Viet Nam	Air-conditioner	Motorbike	Car

\* Note by Turkey: The information in this document with reference to "Cyprus" relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey shall preserve its position concerning the "Cyprus issue".

Note by all the European Union Member States of the OECD and the European Union: The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.



## ANNEX F – TECHNICAL STANDARDS FOR PISA 2015

### INTRODUCTION

The purpose of this annex is to list the set of standards upon which the PISA 2015 data collection activities will be based, as was the case for previous PISA assessments [doc. ref. EDU/PISA/GB(2009)17/REV1]. In following the procedures specified in the standards, the partners involved in the data collection activities contribute to creating an international dataset of a quality that allows for valid cross-national inferences to be made.

The standards for data collection and submission were developed with three major, and inter-related, goals in mind: consistency, precision and generalisability of the data. Furthermore, the standards serve to ensure a timely progression of the project in general.

- **Consistency:** Data should be collected in an equivalent fashion in all countries, using equivalent test materials. A comparable sample of the student population should perform under test conditions that are as similar as possible. Given consistent data collection (and sufficiently high response rates), test results are likely to be comparable across regions and countries. The test results in different countries will reflect differences in the performance of the students measured, and will not be caused by factors which are un-related to performance.
- **Precision:** Data collection and submission practices should leave as little room as possible for spurious variation or error. This holds for both systematic and random error sources, e.g. when the testing environment differs from one group of students to another, or when data entry procedures leave room for interpretation. An increase in precision relates directly to the quality of results one can expect: The more precise the data, the more powerful the (statistical) analyses, and the more trustworthy the results to be obtained.
- **Generalisability:** Data are collected from specific individuals, in a specific situation, and at a certain point in time. Individuals to be tested, test materials and tasks etc. should be selected in a way that will ensure that the conclusions reached from a given set of data do not simply reflect the setting in which the data were collected but hold for a variety of settings and are valid in the target population at large. Thus, collecting data from a representative sample of the population, for example, will lead to results that accurately reflect the level of literacy of fifteen-year-old students in a country.
- **Timeliness:** Consistency, precision and generalisability of the data can be obtained in a variety of ways. However, the tight timelines and budgets in PISA, as well as the sheer number of participating countries, preclude the option of developing and monitoring local solutions to be harmonized at a later stage in the project. Therefore, the standards specify one clear-cut path along which data collection and data submission should progress.

This document strives to establish a collective agreement of mutual accountability among countries, and of the International Contractor towards the countries. This document details each standard, its rationale, and the quality assurance data that need to be collected to demonstrate that the standard has been met.

Where standards have been fully met, data will be recommended for inclusion in the PISA 2015 dataset. Where standards have not been fully met, an adjudication process will determine the extent to which the quality and international comparability of the data have been affected. The result of data adjudication will determine whether the data will be recommended for inclusion in the PISA2015 dataset.

Since attaining the various standards is cumulative and potentially interactive (i.e. .not attaining standard X is NOT the same as not attaining standards X, Y and Z), in principle each dataset should be evaluated against all standards jointly. Also, it is possible that countries' proposed plans for implementation are not, for various and often unforeseen circumstances, actually implemented (e.g. national teacher strike affecting not only response rates but also testing conditions; unforeseen National Centre budget cuts which impact on print and data management quality). Therefore, the final evaluation of standards needs to be made with respect to the data as submitted since this is the definitive indication of what may appear in the released international dataset.

If any issues with attaining standards are identified, the International Project Director initiates communication with the National Centre as soon as possible. Priority in communication rectifies the identified issues.

The PISA standards act as a benchmark of best practice. As such, the standards are designed to assist national centres and International Contractors by explicitly indicating the expectations of data quality and study implementation endorsed by the PISA Governing Board, and by clarifying the timelines of the activities involved. The standards formulate levels of attainment, while timelines and feedback schedules of both the participating countries and the International Contractors are defined in the *PISA operations manuals*.



As specified in the Contracts for the Implementation of the sixth cycle of the OECD Programme for International Student Assessment, the International Contractor for Core 4 takes responsibility for developing and implementing procedures for assuring data quality. Therefore, the International Contractor for Core 4 mediates, and monitors the countries' activities specified in this document, while the adherence to the standards by all International Contractors is monitored by the participating countries via the OECD Secretariat.

Where the technical standards stipulate that variations from the standards require agreement between participating countries and the International Contractors, National Project Managers are asked to initiate the process of negotiation and to undertake everything possible to facilitate an agreement. Where agreement between National Project Managers and the International Contractors cannot be reached, the OECD will adjudicate and resolve the issues. The OECD will also adjudicate any issues resulting from non-compliance with the technical standards that cannot be resolved between participating countries and the Contractors.

There are three types of standards in this document; each with a specific purpose:

- **Data Standards** refer to aspects of study implementation that directly concern the quality of the data or the assurance of that quality. These standards have been endorsed by the Technical Advisory Group and wherever proportions or quantities are specified (for example, response rates), these have been reached through examination of research undertaken or have been reviewed by members of the Technical Advisory Group with the aim of minimising the effect of any potential bias in the data.
- **Management Standards** are in place to ensure that all PISA operational objectives are met in a timely and coordinated manner.
- **National Involvement Standards** reflect the expectations set out in the PISA 2015 Terms of Reference that the content of the PISA tests is established in consultation with national representatives with international content expertise. In particular, these standards ensure that the internationally developed instruments are widely examined for cross-national, cross-cultural and cross-linguistic validity and that the interests and involvement of national stakeholders are considered throughout the study.

## FORMAT OF THE DOCUMENT

The standards are grouped into sections that relate to specific tasks in the PISA data collection process. For every section, a rationale is given explaining why standard setting is necessary. The standards in each section consist of three distinct elements. First, there are the **Standards** themselves that are numbered and are shown in shaded boxes. Second, there are **Notes** that provide additional information on the standards directly. The notes are listed after the standards in each section. Third, there are the **Quality Assurance** measures that will be used to assess if a standard has been met or not. These are listed at the end of each section. In addition, the standards contain words that have a defined meaning in the context of the standards. These words are shown in *italics* throughout the document and are clarified in the **Definitions** section at the end of the document, where the terms are listed alphabetically.

## SCOPE

The standards in this document apply to data from *adjudicated entities* that include both *PISA participants* and *additional adjudicated entities*. The PISA Governing Board will approve the list of adjudicated entities to be included in a PISA cycle.

## DATA STANDARDS

### 1. Target population and sampling

**Rationale:** Meeting the standards specified in this section will ensure that in all countries, the students tested come from the same target population in every country, and are in a nearly equivalent age range. Therefore, the results obtained will not be confounded by potential age effects. Furthermore, to be able to draw conclusions that are valid for the entire population of fifteen-year-old students, a representative sample shall be selected for participation in the test. The size of this representative sample should not be too small, in order to achieve a certain precision of measurement in all countries. For this reason, minimum numbers of participating students and schools are specified.

The mode of drawing the samples used in the study is crucial to data quality. The goal of the project is to collect data that are representative for the population at large, in such a way that the reliability of the results can be quantified. To reach this goal the sampling procedures must follow established scientific principles for drawing samples from finite populations.



<b>Standard 1.1</b>	The PISA <i>Desired Target Population</i> is agreed upon through negotiation between the National Project Manager and the International Contractor for Core 5, within the constraints imposed by the definition of the <i>PISA Target Population</i> .
<b>Standard 1.2</b>	Unless otherwise agreed upon only <i>PISA-Eligible students</i> participate in the test.
<b>Standard 1.3</b>	Unless otherwise agreed upon, the <i>testing period</i> :
	<ul style="list-style-type: none"> <li>▪ is no longer than six consecutive weeks in duration;</li> <li>▪ does not coincide with the first six weeks of the academic year; and</li> <li>▪ begins exactly three years from the beginning of the <i>testing period</i> in the previous PISA cycle, unless otherwise agreed upon.</li> </ul>
<b>Standard 1.4</b>	Schools are sampled using <i>agreed upon</i> , established and professionally recognised principles of scientific sampling.
<b>Standard 1.5</b>	Student lists should not be collected more than 8 weeks prior to the start of data collection, unless otherwise agreed upon.
<b>Standard 1.6</b>	Students are sampled using <i>agreed upon</i> , established and professionally recognised principles of scientific sampling and in a way that represents the full population of <i>PISA-Eligible students</i> .
<b>Standard 1.7</b>	The PISA Defined Target Population covers 95% or more of the <i>PISA Desired Target Population</i> . That is, <i>school-level exclusions</i> and <i>within-school exclusions</i> combined do not exceed 5%.
<b>Standard 1.8</b>	The student sample size for the <b>computer-based mode including Collaborative Problem Solving</b> is a minimum of 5,400 assessed students for <i>PISA participants</i> and 1,800 assessed students for <i>additional adjudicated entities</i> , or the entire <i>PISA Defined Target Population</i> where the <i>PISA Defined Target Population</i> is below 5,400 and 1,800 respectively. The student sample size for the <b>paper-based mode</b> or the <b>computer-based mode without Collaborative Problem Solving</b> is a minimum of 4,500 assessed students for <i>PISA participants</i> and 1,500 assessed students for <i>additional adjudicated entities</i> , or the entire <i>PISA Defined Target Population</i> where the <i>PISA Defined Target Population</i> is below 4,500 and 1,500 respectively.
<b>Standard 1.9</b>	The school sample size is a minimum of 150 schools for <i>PISA participants</i> , and 50 schools for <i>additional adjudicated entities</i> , or all schools that have students in the <i>PISA Defined Target Population</i> where the number of schools with students in the <i>PISA Defined Target Population</i> is below 150 and 50 respectively.
<b>Standard 1.10</b>	The final weighted school response rate is at least 85% of sampled schools. If a response rate is below 85% then an acceptable response rate can still be achieved through <i>agreed upon</i> use of replacement schools.
<b>Standard 1.11</b>	The final weighted student response rate is at least 80% of all sampled students across responding schools.
<b>Standard 1.12</b>	The final weighted sampling unit response rate for any International Option which requires response rates, is at least 80% of all sampled units across responding International Option schools.
<b>Standard 1.13</b>	Unless otherwise agreed upon, the International Contractor for Core 5 will draw the school sample for the Main Survey
<b>Standard 1.14</b>	Unless otherwise agreed upon, the National Centre will use KeyQuest to draw the student sample, using the list of eligible students provided for each school.

**Note 1.1** The Target Population and Sampling standard apply to the Main Survey but not the Field Trial.

**Note 1.2** Data from schools where the student response rate is greater than 25% will be included in the PISA dataset.

**Note 1.3** For the purpose of calculating school response rates, a participating school is defined as a sampled school in which more than 50% of sampled students respond.

**Note 1.4** Guidelines for acceptable exclusions that do not affect standard adherence, are as follows:

- *School level exclusions* that are exclusions due to geographical inaccessibility, extremely small school size, administration of PISA would be not feasible within the school, and other *agreed upon* reasons and that total to less than 0.5% of the *PISA Desired Target Population*;
- *School level exclusions* that are due to a school containing only students that would be *within-school exclusions* and that total to less than 2.0% of the *PISA Desired Target Population*; and
- *Within-school exclusions* that total to less than 2.5% of the *PISA Desired Target Population* – these exclusions could include, for example, students not able to do the test because of a functional disability.

**Note 1.5** Principles of scientific sampling include, but are not limited to:

- The identification of appropriate stratification variables to reduce sampling variance and facilitate the computation of non-response adjustments.
- The incorporation of an agreed target cluster size of PISA-Eligible students. For computer-based assessment, the target cluster size is 42 students. For paper-based assessment, or computer-based without collaborative problem solving, the target cluster size is 35. *Upon agreement* this can be increased, or reduced to a number not less than 20.

**Note 1.6** Any exceptional costs associated with verifying a school sample taken by the National Centre, or a student sample selected other than by using KeyQuest will be borne by the National Centre.

**Note 1.7** Agreement with the International Contractor of alternative methods of drawing samples will be subject to the principle that the sampling methods used are scientifically valid and consistent with PISA's documented sampling methods. Where a PISA participating country chooses to draw the school sample, the National Centre provides the International Contractor with the data and documentation required for it to verify the correctness of the sampling procedures applied. Where a PISA participating country chooses not to use KeyQuest to draw the student sample, the National Centre provides the International Contractor with the data and documentation required for it to verify the correctness of the sampling procedures applied.



## Quality assurance

- Sampling procedures as specified in the *PISA operations manuals*.
- School sample drawn by International Contractor for Core 5 (or if drawn by the national centre, then verified by the International Contractor for Core 5).
- Student sample drawn through KeyQuest (or if drawn by other means, then verified by the International Contractor for Core 5).
- Sampling forms submitted to the International Contractor for Core 5.
- Main Survey Review Quality Assurance Survey.

## 2. Language of testing

**Rationale:** Using the language of instruction will ensure analogous testing conditions for all students within a country, thereby strengthening the consistency of the data. It is assumed that the students tested have reached a level of understanding in the language of instruction that is sufficient to be able to work on the PISA test without encountering linguistic problems (see also the criteria for excluding students from the potential assessment due to insufficient experience in the language of assessment: *within-school exclusions*). Thus, the level of literacy in reading, mathematics and science can be assessed without interference due to a critical variation in language proficiency.

<b>Standard 2.1</b>	<p>The PISA test is administered to a student in a language of instruction provided by the sampled school to that sampled student in the major domain (Science) of the test.</p> <p>If the language of instruction in the major domain is not well defined across the set of sampled students then, if <i>agreed upon</i>, a choice of language can be provided, with the decision being made at the student, school, or National Centre level. Agreement with the International Contractor will be subject to the principle that the language options provided should be languages that are common in the community and are common languages of instruction in schools in that <i>adjudicated entity</i>.</p> <p>If the language of instruction differs across domains then, if <i>agreed upon</i>, students may be tested using assessment instruments in more than one language on the condition that the test language of each domain matches the language of instruction for that domain. Information obtained from the Field Trial will be used to gauge the suitability of using assessment instruments with more than one language in the Main Survey.</p> <p>In all cases the choice of test language(s) in the assessment instruments is made prior to the administration of the test.</p>
---------------------	--

## 3. Field trial participation

**Rationale:** The Field Trial gives countries the opportunity to try out the logistics of their test procedures and allows the International Contractors to make detailed analyses of the items so that only suitable ones are included in the Main Survey.

<b>Standard 3.1</b>	<p><i>PISA participants</i> participating in the PISA 2015 Main Survey will have successfully implemented the Field Trial. Unless otherwise agreed upon:</p> <ul style="list-style-type: none"> <li>▪ A Field Trial should occur in an assessment language if that language group represents more than 5% of the target population.</li> <li>▪ For assessment languages that apply to between 5 and 50% of the target population, the Field Trial student sample should be a minimum of 100 students per item.</li> <li>▪ For languages that apply to more than 50% of the target population, the Field Trial student sample should be a minimum of 200 students per item.</li> <li>▪ For additional adjudicated entities, where the assessment language applies to between 5 and 100% of the target population in the entity, the Field Trial student sample should be a minimum of 100 students per item.</li> </ul>
<b>Standard 3.2</b>	Countries planning to use computer-based delivery in 2015 must also field trial paper-and-pencil booklets to test for mode effects.

**Note 3.1** The PISA Technical Standards for the Main Survey generally apply to the Field Trial, except for the Target Population standard, the Sampling standard, and the Quality Monitoring standard. For the Field Trial a sampling plan needs to be *agreed upon*.

**Note 3.2** The Field Trial participation standard for assessment languages applicable to between 5 and 50% of the target population can be varied if *agreed upon*, with such agreement subject to the principle that the absence of a Field Trial for that language would not affect the Main Survey and the principle that the assessment language version is trialled in another *adjudicated entity* where the assessment language applies to more than 50% of the target population.

**Note 3.3** The sample size for the Field Trial will be a function of the test design and will be set to achieve the standard of 200 student responses per item.

**Note 3.4** Consideration will be given to reducing the required number of students per item in the Field Trial where there are fewer than 200 students in total expected to be assessed in that language in the Main Survey.

**Note 3.5** Without testing for mode effects in the field trial, it will be impossible for countries who wish to deliver PISA 2015 on computer to measure trends relative to performance in previous paper-based cycles.



## 4. Adaptation of tests, questionnaires and manuals

**Rationale:** In order to be able to assess how the performance in a country has evolved from one PISA cycle to the other, the same instruments have to be used in all assessments. If instruments differ, then it is unclear whether changes in performance reflect changes in literacy or whether they just mirror the variation in the test items. The same holds true for the assessment instruments that are used within a PISA cycle: To validly compare performance across countries, all assessment instruments have to be as similar as possible. In fact, it is of utmost importance to provide equivalent information for the students in all countries that take part in the study. Therefore, not only the assessment instruments, but also the instructions given to the students, and the procedures of data-collection have to be equivalent. To achieve this goal, other individuals who play a key role in the data-collection process, i.e. the test administrators, school coordinators, and school associates, should receive the same information in all participating countries.

<b>Standard 4.1</b>	The majority of test items used for linking are administered unchanged from their previous administration. The computer-based versions will include instructions as to the appropriate response mode for each item and may require some minor revision as noted in 4.2 below.
<b>Standard 4.2</b>	All assessment instruments are psychometrically equivalent to the <i>source versions</i> . <i>Agreed upon</i> adaptations to the local context are made if needed.
<b>Standard 4.3</b>	National versions of questionnaire items used in previous cycles will be administered unchanged from their previous administration, unless amendments have been made to <i>source versions</i> .
<b>Standard 4.4</b>	The questionnaire instruments are equivalent to the <i>source versions</i> . <i>Agreed upon</i> adaptations to the local context are made if needed.
<b>Standard 4.5</b>	The Test Administrator Manual and the School Coordinator Manual (or the School Associate Manual) are equivalent to the <i>source versions</i> . <i>Agreed upon</i> adaptations to the local context are made if needed.

**Note 4.1** The quality assurance requirements for this standard apply to instruments that are in an assessment language used as a language of instruction for more than 5% of the target population.

**Note 4.2** In a very few cases, stimulus materials will be adjusted so they can be presented consistently across countries on the computer screen. The Field Trial mode study will be used to investigate whether such changes impact item performance.

### Quality assurance

- *Agreed Upon* Manual Adaptation Spreadsheet (MAS) and Questionnaire Adaptation Spreadsheet (QAS).
- Test Adaptation Spreadsheet (TAS), Booklet Adaptation Spreadsheet (BAS), and Computer-Based Assessment Adaptation Forms in which adaptations to assessment units, common booklet parts and coding guides are documented. Adaptations will be checked for compliance with the PISA Translation and Adaptation Guidelines by international verifiers, and the verifiers' recommendations will be vetted by the translation referee.
- Verifier Reports (statistics generated by the TAS and computer-based assessment adaptation forms, in combination with a short qualitative report).
- Final Check Report, including check of interventions that require follow-up.
- Field Trial and Main Survey Review Quality Assurance Surveys.
- Item and scale statistics generated by the International Contractors for Core 3 (assessment materials) and Core 6 (questionnaires).

## 5. Translation of assessment instruments, questionnaires and manuals

**Rationale:** To be able to compare the performance of students across countries, and of students with different instruction languages within a country, the linguistic equivalence of all materials is central. While Standards 4.1 to 4.4 serve to ensure that equivalent information is given to the students in all countries involved, in general, the following Standards 5.1 and 5.2 emphasise the importance of language. Again the goal is to ensure that literacy will be assessed, and not variations of information caused by differences in the translation of materials.



<b>Standard 5.1</b>	The following documents are translated into the assessment language in order to be linguistically equivalent to the international <i>source versions</i> . <ul style="list-style-type: none"> <li>▪ All administered assessment instruments</li> <li>▪ All administered questionnaires</li> <li>▪ The Test Administrator script from the Test Administrator (or School Associate) Manual</li> <li>▪ The Coding Guides</li> </ul>
<b>Standard 5.2</b>	Unless otherwise <i>agreed upon</i> , the following documents are translated/adapted into the assessment language to make them linguistically equivalent to the international <i>source versions</i> . <ul style="list-style-type: none"> <li>▪ The Test Administrator Manual</li> <li>▪ The School Coordinator Manual</li> </ul> OR <ul style="list-style-type: none"> <li>▪ The School Associate Manual (in the case of countries using School Associates)</li> </ul> In the case of the manuals, only <i>specified parts</i> are made linguistically equivalent.

**Note 5.1** The quality assurance requirements for this standard apply to instruments that are in a language that is administered to more than 10% of the target population.

**Note 5.2** The “specified parts” of manuals will be described in national centre operational manuals.

### **Quality assurance**

- *Agreed upon Translation Plan* developed in accordance with the specifications in the PISA operations manuals where the *Translation Plan* would require double translation by independent translators.
- *Agreed Upon Questionnaire Adaptation Spreadsheet (QAS)*
- Test Adaptation Spreadsheet (TAS), Booklet Adaptation Spreadsheet (BAS) and computer-based assessment adaptation forms in which adaptations to assessment units, common booklet parts and coding guides are documented. Adaptations will be checked for compliance with the PISA Translation and Adaptation Guidelines by international verifiers, and the verifiers’ recommendations will be vetted by the translation referee.
- Verifier Reports (statistics generated by the TAS and computer-based assessment adaptation forms, in combination with a short qualitative report)
- Final Check report (test booklets and questionnaires only)
- Submitted test booklets and computer-based assessments as used in the study
- Field Trial and Main Survey Review Quality Assurance Surveys
- Item and scale statistics generated by the International Contractors for Core 3 (assessment materials) and Core 6 (questionnaires)

## **6. Testing of national software versions**

Rationale: Countries must thoroughly test and validate the national software releases that are used to deliver the PISA computer-based instruments in schools, as well as the online questionnaires that are delivered via the Internet.

<b>Standard 6.1</b>	The International Contractors must test all national software versions prior to their release to ensure that they were assembled correctly and have no technical problems.
<b>Standard 6.2</b>	Once released, countries must test the national software versions following testing plans to ensure the correct implementation of national adaptations and extensions, display of national languages, and proper functioning on computers typically found in schools in each country.

**Note 6.1** Errors found during testing should be promptly communicated to the International Contractors using agreed-upon problem reporting procedures. These procedures require that testing results are shared with the International Contractors in order to monitor the quality of the instruments.



### **Quality assurance**

- Detailed testing plans
- Review of testing results

## **7. Technical support**

Rationale: Countries participating in the computer-based delivery mode will be primarily responsible for resolving PISA-related operational issues in their countries, including hardware issues and provision of technical support to schools and test administrators.

**Standard 7.1** Each country should have a designated PISA helpdesk with contact information provided to each of its test administrators and school coordinators.

**Standard 7.2** The country helpdesk staff must:

- be familiar with the PISA computer system requirements applications and training materials,
- be familiar with all national software standards and procedures; and
- attend the test administrator training sessions to become familiar with the computer-based assessments and appreciate the challenges faced by schools and test administrators.

### **Quality assurance**

- National Centre Quality Monitoring
- Field Trial and Main Survey Review Quality Assurance Surveys

## **8. Test administration**

Rationale: Certain variations in the testing procedure are particularly likely to affect test performance. Among them are session timing, the administration of test materials and support material like rulers and calculators, the instructions given prior to testing, the rules for excluding students from the assessment etc. A full list of relevant test conditions is given in the *PISA operations manuals*. To ensure that the data are collected consistently, and in a comparable fashion, for all participants, it is therefore very important to keep the chain of action in the data-collection process as constant as possible.

Furthermore, the goal of the assessment is to arrive at results which cover a wide range of areas. Given the time constraints, any one student is presented only with a certain portion of the test items. Moreover, to preclude sources of random error unforeseen by the test administrators and the test designers, the students taking part in the survey have to be selected *a-priori*, in a statistically random fashion. Only then will the students participating in the study mirror the population of fifteen-year-old students in the country. The statistical analysis will take this sampling design into account, thereby arriving at results that are representative for the population at large. For these reasons, it is of utmost importance to assign the proper test booklets to the participants specified beforehand. The student tracking form is central in monitoring whether this goal has been achieved.

The test administrator plays a central role in all of these issues. Special consideration is therefore given to the training of the test administrators, ensuring that as little variation in the data as possible is caused by random or systematic variation in the activities of test administrators.

An important part of the testing situation relates to the relationship between test administrators and test participants. Therefore, any personal interaction between test administrators and students, either in the past or in the testing situation, counteracts the goal of collecting data in a consistent fashion across countries and participants. Strict objectivity of the test administrator, on the other hand, is instrumental in collecting data that reflect the level of literacy obtained, and that are not influenced by factors un-related to literacy. The results based on these data will be representative for the population under consideration.



<b>Standard 8.1</b>	All test sessions follow international procedures as specified in the PISA operations manuals, particularly the procedures that are: <ul style="list-style-type: none"> <li>▪ relating to test session timing,</li> <li>▪ for maintaining test conditions,</li> <li>▪ for student tracking, and</li> <li>▪ for assigning assessment materials.</li> </ul>
<b>Standard 8.2</b>	The relationship between Test Administrators and participating students must not compromise the credibility of the test session. In particular, the Test Administrator should not be the reading, mathematics, or science instructor of any student in the assessment sessions he or she will administer for PISA.

**Note 8.1** Test Administrators should preferably not be school staff.

#### **Quality assurance**

- Test Administrator's Test Session Report Forms
- PISA Quality Monitors
- Main Survey Review Quality Assurance Survey

### **9. Training support**

Rationale: NPMs or their designees shall participate in a train-the-trainer session conducted by qualified contractor staff. This ensures standardisation of training delivery to test administrators, allows trainers to become familiar with PISA materials and procedures, and informs trainers of their responsibilities for overseeing the PISA testing.

<b>Standard 9.1</b>	Qualified contractor staff will conduct trainer training sessions with NPMs or designees on PISA materials and procedures to prepare them to train PISA test administrators.
<b>Standard 9.2</b>	NPMs or designees shall use the comprehensive training package developed by the contractors to train PISA test administrators.
<b>Standard 9.3</b>	All test administrator training sessions should be scripted to ensure consistency of presentations across training sessions and across countries. Failure to do so could cause errors in data collection and invalidate the results.
<b>Standard 9.4</b>	In-person test administrator trainings should be conducted by the NPMs or designees, unless a suitable alternative is agreed upon.
<b>Standard 9.5</b>	PQMs need to successfully complete self-training materials and attend webinars to review and enhance the self-training.

**Note 9.1** Test administrator refers to any person officially assigned to conduct a PISA testing session.

#### **Quality assurance**

- Participation in trainer training sessions in standardised procedures by qualified contractor staff
- National Centre Quality Monitoring
- Field Trial and Main Survey Review Quality Assurance Surveys
- Monitored training modules of PQMs

### **10. Implementation of national options**

Rationale: These standards serve to ensure that for students participating both in the international and the national survey, the national instruments will not affect the data used for the international comparisons. Data are therefore collected consistently across countries, and potential effects like test fatigue, or learning effects from national test items, are precluded.



**Standard 10.1** Only *national options* that are *agreed upon* between the National Centre and the International Contractors are implemented.

**Standard 10.2** Any *national option* instruments that are not part of the core component of PISA are administered after all the test and questionnaire instruments of the core component of PISA have been administered to students that are part of the international PISA sample.

## 11. Security of the material

Rationale: The goal of the PISA assessment is to measure the literacy levels in the content domains. Prior familiarisation with the test materials, or training to the test, will heavily degrade the consistency and validity of the data. In the extreme case, the results would only reflect how well participants are able to memorise the test items. In order to be able to assess the competencies obtained during schooling rather than short-term learning success, and to make valid international comparisons, confidentiality is extremely important.

**Standard 11.1** PISA materials designated as secure are kept confidential at all times. Secure materials include all test materials, data, and draft materials. In particular:

- no-one other than approved project staff and participating students during the test session is able to access and view the test material,
- no-one other than approved project staff will have access to secure PISA data and embargoed material, and
- formal confidentiality arrangements will be in place for all approved project staff.

### Quality assurance

- Security arrangements as specified in the *PISA operations manuals* or *agreed upon* variation
- National Centre Quality Monitoring
- Field Trial and Main Survey Review Quality Assurance Surveys

## 12. Quality monitoring

Rationale: To obtain valid results from the assessment, the data collected have to be of high quality, i.e. they have to be collected in a consistent, reliable and valid fashion. This goal is implemented first and foremost by the test administrators, who are seconded by the quality monitors. The quality monitors provide country-wide supervision of all data-collection activities.

**Standard 12.1** PISA test administration is monitored using site visits by trained independent quality monitors.

**Standard 12.2** An agreed number of site visits to observe test administration sessions are conducted in each PISA participating country/economy.

**Standard 12.3** Test administration sessions that are the subject of a site visit are selected by the International Contractor for Core 4 to be representative of a variety of schools in a country/economy.

**Note 12.1** A failure to meet the Quality Monitoring standard in the Main Survey will lead to a significant lack of quality assurance data for other standards.

**Note 12.2** The Quality Monitoring standards apply to the Main Survey but not to the Field Trial.

**Note 12.3** The National Centre provides the International Contractor for Core 4 the assistance required to implement the site visits effectively.

### Quality assurance

- Curricula Vitae of the PISA Quality Monitor nominees forwarded by the National Project Manager to the International Contractor for Core 4.
- PISA Quality Monitor Reports
- National Centre Quality Monitor Report



## 13. Printing of material

**Rationale:** Variations in print quality may affect data quality. When the quality of paper and print is very poor, the performance of students is influenced not only by their levels of literacy, but also by the degree to which test materials are legible. To rule out this potential source of error, and to increase the consistency and precision of the data collection, paper and print quality samples are solicited from national centres in their first cycle of participation.

- Standard 13.1** All student assessment material is printed using an agreed upon paper and print quality.
- Standard 13.2** The cover page of all PISA assessment instruments used in schools contains all information as specified by the PISA Governing Board.
- Standard 13.3** The layout and pagination of all test material is the same as in the *source versions*, unless otherwise agreed upon.
- Standard 13.4** The layout and formatting of the questionnaire material is equivalent to the *source versions*.

**Note 13.1** For National Centres that have participated in previous cycles, PISA instruments used in previous cycles or from the Field Trial preceding the Main Survey that have been submitted to the previous International Contractor can be used for the purpose of agreeing on printing quality where the national centre indicates that printing and paper of the same standard will be used. Otherwise, National Centres will submit a sample of printed material to the International Contractor for Core 4 for agreement, including the cover and selected items as specified in the *PISA operations manuals*.

**Note 13.2** The cover page of all PISA assessment instruments used in schools should contain all information necessary to identify the material as being part of the data-collection process for PISA, and for checking whether the data collection follows the assessment design, i.e. whether the mapping of the student on the one hand, and test booklets and questionnaires, on the other, have been correctly established. The features of the cover page referred to in Standard 13.2 are specified in the *PISA operations manuals*.

### Quality assurance

- Submitted sample or agreement that quality will be similar to previous cycle or Field Trial versions.
- Materials submitted to the International Contractor for Core 4, as described in note 13.1 above.
- Field Trial and Main Survey Review Quality Assurance Surveys.

## 14. Response coding<sup>1</sup>

**Rationale:** To ensure the comparability of the data, the responses from all test participants in all participating countries have to be coded following one single coding scheme. Therefore, all coding procedures have to be standardised, and coders have to complete training sessions to master this task.

- Standard 14.1** The coding scheme described in the coding guide in the distributed items is implemented according to instructions from the International Contractor's item developers.
- Standard 14.2** Representatives from each National Centre attend the international PISA coder training session for both the Field Trial and the Main Survey.
- Standard 14.3** Both the single and multiple coding procedures as specified in the *PISA operations manuals* (see Note 14.1), or an agreed upon variation thereof, are implemented.
- Standard 14.4** Coders are recruited and trained following agreed procedures.

**Note 14.1** Preferred procedures for recruiting and training coders are outlined in the *PISA operations manuals*.

**Note 14.2** The optimum number of Coder Training session participants would depend on factors such as the expertise of National Centre staff, and resource availability.

### Quality assurance

- Indices of inter-coder agreement
- Field Trial and Main Survey Review Quality Assurance Surveys
- .....

1. The terms coding, coders and codes are used instead of other terms such as marking, markers, marks, rating and raters.



## 15. Data submission

**Rationale:** The timely progression of the project, within the tight timelines given depends on the quick and efficient submission of all collected data. Therefore, one single data submission format is proposed, and countries are asked to submit only one database to the International Contractor for Core 3. Furthermore, to avoid potential errors when consolidating the national databases, any changes in format that were implemented subsequent to the general agreement have to be announced.

**Standard 15.1** Each *PISA* participant submits its data in a single complete database, unless otherwise agreed upon.

**Standard 15.2** All data collected for *PISA* will be imported into a national database using the Data Management Expert (DME) data integration software provided by the International Contractor for Core 3 following specifications in the corresponding operational manuals and international/national record layouts (codebooks). Data are submitted in the DME format.

**Standard 15.3** Data for all instruments are submitted. This includes the assessment data, questionnaire data, and tracking data as described in the *PISA operations manuals*.

**Standard 15.4** Unless agreed upon, all data are submitted without recoding any of the original response variables.

**Standard 15.5** Each *PISA* participating country's database is submitted with full documentation as specified in the *PISA operations manuals*.

## MANAGEMENT STANDARDS

### 16. Communication with the international contractors

**Rationale:** Given the tight schedule of the project, delays in communication between the National Centres and the International Contractors should be minimised. Therefore, National Centres need continuous access to the resources provided by the International Contractors.

**Standard 16.1** The International Contractors ensure that qualified staff are available to respond to requests by the National Centres during all stages of the project. The qualified staff:

- are authorised to respond to National Centre queries,
- acknowledge receipt of National Centre queries within one working day,
- respond to coder queries from National Centres within one working day,
- respond to other queries from National Centres within five working days, or, if processing the query takes longer, give an indication of the amount of time required to respond to the query.

**Standard 16.2** The National Centre ensures that qualified staff are available to respond to requests by the International Contractors during all stages of the project. The qualified staff:

- Are authorised to respond to queries,
- Acknowledge receipt of queries within one working day,
- Respond to queries from International Contractors within five working days, or, if processing the query takes longer, give an indication of the amount of time required to respond to the query.

**Note 16.1** Response timelines and feedback schedules for the National Centres and the International Contractor are further specified in the *PISA operations manuals*.



## 17. Notification of international and national options

Rationale: Given the tight timelines, the deadlines given in the following two standards will enable the International Contractor to progress with work on time.

- Standard 17.1** *National options* are agreed upon before 1 December in the year preceding the Field Trial and before 1 December in the year preceding the Main Survey.
- Standard 17.2** The national centre notifies the International Contractor of its intention to participate in specific international options three months prior to the start of the translation period.

## 18. Schedule for submission of materials

Rationale: To meet the requirements of the work programme, and to progress according to the timelines of the project, the International Contractor will need to receive a number of materials on time.

- Standard 18.1** An agreed upon *Translation Plan* will be negotiated between each national centre and the International Contractors.
- Standard 18.2** The following items are submitted to the International Contractors in accordance with *agreed timelines*:
  - the Translation Plan
  - a print sample of booklets prior to final printing, for countries using the paper-based instruments (where this is required, see Standard 13.1 and Note 13.1)
  - results from the national checking of adapted computer-based assessment materials and questionnaires,
  - sampling forms (see Standard 1)
  - demographic Tables
  - Field Trial and Main Survey Reviews
  - other documents as specified in the PISA operations manuals.
- Standard 18.3** Questionnaire materials are submitted for linguistic verification only after all adaptations have been agreed upon.
- Standard 18.4** All adaptations to those elements of the Test Administrator and School Co-ordinator (or School Associate) manuals that are required to be linguistically equivalent to the source as specified in Standard 5.2, need to be agreed upon.

### Quality assurance

- Agreed upon Translation Plan
- International Contractors' records
- Assessment materials submitted for linguistic verification with corresponding adaptation spreadsheets filled in by the National Centre

## 19. Management of data

Rationale: Consolidating and merging the national databases is a time-consuming and difficult task. To ensure the timely and efficient progress of the project, the International Contractors need continuous access to national resources helping to rule out uncertainties and to resolve discrepancies. This standard aims to prevent substantial delays to the whole project which could result from a delay in processing the data of a small number of participating countries.



- Standard 19.1** The timeline for submission of national databases to the International Contractors is within eight weeks of the last day of testing for the Field Trial and within twelve weeks of the last day of testing for the Main Survey, unless otherwise *agreed upon*.
- Standard 19.2** National Centres execute data checking procedures as specified in the *PISA Operation Manuals* before submitting the database.
- Standard 19.3** National Centres make a data manager available upon submission of the database. The data manager:
- is authorised to respond to International Contractor data queries
  - is available for a three-month period immediately after the database is submitted unless otherwise *agreed upon*
  - is able to respond to International Contractor queries within three working days
  - is able to resolve data discrepancies.
- Standard 19.4** A complete set of PISA paper-based instruments as administered and including any *national options*, is forwarded to the International Contractor for Core 4 on or before the first day of testing. The submission includes the following:
- hard copies of instruments
  - electronic PDF copies of instruments.
- Standard 19.5** To enable the *PISA participant* to submit a single dataset, all instruments for all *additional adjudicated entities* will contain the same variables as the primary *adjudicated entity* of the *PISA participant*.

**Note 19.1** Each participating country/economy will receive its own national micro-level PISA database (the “national database”), in electronic form as soon as it has been processed from the International Contractors for PISA. The national database will contain the complete set of responses from the students, parents, school principals and surveyed participants in that country/economy.

Each participating country/economy has access to and can publish its own data **after** a date that is established by the PISA Governing Board for the publication of the initial OECD publication of the survey results (the “initial international OECD publication”).

The OECD Secretariat will not release national data to other countries/economies until participating countries/economies have been given an opportunity to review and comment on their own national data and until the release of such data has been approved by the national authorities.

A deadline and procedures for withdrawing countries/economies’ national data from the international micro-level PISA database (the “international database”) will be decided upon by the PISA Governing Board. Countries/economies can withdraw data only prior to obtaining access to data from other countries/economies. Withdrawn data will not be made available to other countries/economies.

The PISA Governing Board will discuss with participating countries/economies whose data manifests technical anomalies as to whether the data concerned can be included in the international database. The decision of the PISA Governing Board will be final. Participating countries/economies may, however, continue to use data that are excluded from the international database at the national level.

The Contractor for Core 3 will then compile the international database, which will comprise the complete set of national PISA databases, except those data elements that have been withdrawn by participating countries/economies or by the PISA Governing Board at the previous stage. The international database will remain confidential until the date on which the initial international OECD publication is released.

National data from all participating countries/economies represented in the international database will be made available to all participating countries/economies from the date on which the initial international OECD publication is released.

After release of the initial international OECD publication, the international database will be made publicly available on a cost-free basis, through the OECD Secretariat. The database may not be offered for sale.

The international database will form the basis for OECD indicator reports and publications.

The International Contractors for PISA 2015 will have no ownership of instruments or data nor any rights of publication and will be subject to the confidentiality terms set in this agreement.

The OECD establishes rules to ensure adherence to the above procedure and to the continued confidentiality of the PISA data and materials until the agreed release dates. These include confidentiality agreements with all individuals that have access to the PISA material prior to its release.

As guardian of the process and producer of the international database, the OECD will hold copyright in the database and in all original material used to develop, or be included in, the PISA Field Trial and PISA Main Survey (among them the assessment materials, field manuals, and coding guides) in any language and format.

### **Quality assurance**

- International Contractors’ Records

## **20. Archiving of materials**

Rationale: The International Contractors will maintain an electronic archive. This will provide an overview of all materials used and ensure continuity of materials available in participating countries across PISA survey cycles, therefore building upon the knowledge gained nationally in the course of the PISA cycles. This will also ensure that the International Contractors have the relevant materials available during data cleaning, when they are first required.



- Standard 20.1** The International Contractors will maintain a permanent electronic archive of all assessment materials, field manuals and coding guides.
- Standard 20.2** The International Contractors will be responsible for archiving all national versions of computer-based assessment materials.
- Standard 20.3** For paper-based materials, the National Project Manager submits one copy of each of the following translated and adapted Main Survey materials to the International Contractors in the source version software format:
- all administered Test Instruments, including *national options*;
  - all administered Questionnaires, including *national options*;
  - Test Administrator, School Coordinator and School Associate manuals; and
  - Coding Guides.
- Standard 20.2** Unless otherwise requested, National Centres will archive all Field Trial materials until the beginning of the Main Survey, and all Main Survey materials until the publication of the international report. Materials to be archived include:
- all respondents' paper-based test booklets and questionnaires,
  - sampling forms,
  - student lists,
  - student tracking instruments, and
  - all data submitted to the International Contractors.

After completion of a survey the National Centre will transfer this archive to the International Contractor for Core 7 who will compile the national archives from all participants and transfer them to OECD after completion of the Main Study.

## NATIONAL INVOLVEMENT STANDARDS

### 21. National feedback

Rationale: National feedback in areas such as test development is important in maintaining the dynamic and collaborative nature of PISA. National feedback ensures that instruments achieve cross-national, cross-cultural and cross-linguistic validity. It also promotes the inclusion of the interests and involvement of national stakeholders.

- Standard 21.1** National Centres develop appropriate mechanisms in order to promote participation, effective implementation, and dissemination of results amongst all relevant national stakeholders.
- Standard 21.2** National Centres provide feedback to the International Contractors on the development of instruments, domain frameworks, the adaptation of instruments, and other domain-related matters that represent the perspectives of the relevant national stakeholders.

**Note 21.1** As a guideline feedback might be sought from the following relevant stakeholders: policy makers, curriculum developers, domain experts, test developers, linguistic experts and experienced teachers.

#### Quality assurance

- National Centre Quality Monitoring
- Documented strategies
- List of committees and groups
- Membership records of representative groups and/or committees
- Meeting records of representative groups and/or committees



## DEFINITIONS

**Additional Adjudicated Entities** – entities in addition to the first and primary entity managed by a *PISA participant*, where a *PISA participant* manages more than one *adjudicated entity*.

**Adjudicated Entity** – a country, geographic region, or similarly defined population, for which the International Contractors fully implements quality assurance and quality control mechanisms and endorses, or otherwise, the publication of separate PISA results.

**Agreed procedures** – procedures that are specified in the *PISA operations manuals*, or variations that are *agreed upon* between the National Project Manager and the International Contractors.

**Agreed timelines** – timelines that are specified in the *PISA operations manuals*, or variations that are *agreed upon* between the National Project Manager and the International Contractors.

**Agreed upon** – variations and definitions agreed upon between the National Project Manager and the International Contractors

**International Contractors website** – The PISA Portal – PISA 2015 project website – can be accessed through the following address: <http://pisaportal.tudor.lu/portal>. This website contains the *source versions* of instruments, manuals and other documents and information relating to National Centres.

**International Option** – optional additional international instruments or procedures designed and fully supported by the International Contractors.

**KeyQuest** – software developed specifically for the PISA project. The software assists with sampling, student tracking and data submission practices that meet the PISA 2015 technical standards.

**National Centre Quality Monitoring** – the procedures by which Core 4 monitors the quality of all aspects of the implementation of the survey by a National Centre.

**National Option** – A *national option* occurs if:

- i) a National Centre administers any additional instrumentation, for example a test or questionnaire, to schools or students that are part of the PISA international sample. Note that in the case of adding items to the questionnaires, an addition of five or more items to either the school questionnaire or the student questionnaire is regarded as a national option.

OR

- ii) a National Centre administers any PISA international instrumentation to any students or schools that are not part of an international PISA sample (age-based or grade-based) and therefore will not be included in the respective PISA international database.

**PISA Defined Target Population** – all *PISA-Eligible students* in the schools that are listed on the school sampling frame. That is, the *PISA Desired Target Population* minus exclusions.

**PISA Desired Target Population** – the *PISA Target Population* defined for a specific adjudicated entity. It provides the most exhaustive coverage of *PISA-Eligible students* in the *participating economy* as is feasible.

**PISA-Eligible Students** – students who are in the *PISA Target Population*.

**PISA Operations Manuals** – manuals provided by the International Contractors, that is the following:

- National Project Manager's Manual (Core 4),
- Test Administrator Manual (Core 4)
- School Coordinator Manual (Core 4)
- School Associate Manual (Core 4)
- School Sampling Preparations Manual (Core 5)
- Student Sampling Manual (Core 5)



- Data Management Manual (Core 3)
- all other key documents referenced within the National Project Manager's manual.
- The preparation of the *PISA operations manuals* will be carried out by the International Contractors and will describe procedures developed by the International Contractors. The manuals will be prepared following consultation with participating countries/economies, the OECD Secretariat, the Technical Advisory Group and other stakeholders.

**PISA Participant** – an administration centre, commonly called a National Centre that is managed by a person, commonly called a National Project Manager, who is responsible for administering PISA in one or more *adjudicated entities*. The National Project Manager must be authorised to communicate with the International Contractor on all operational matters relating to the *adjudicated entities* for which the National Project Manager is responsible.

**PISA Quality Monitor** – a person nominated by the National Project Manager and employed by the International Contractor for Core 4 to monitor test administration quality in an adjudicated entity.

**PISA Target Population** – students aged between 15 years and 3 (completed) months and 16 years and 2 (completed) months at the beginning of the *testing period*, attending educational institutions located within the *adjudicated entity*, and in grade 7 or higher. The age range of the population may vary up to one month, either older or younger, but the age range must remain 12 months in length. That is, the population can be as young as between 15 years and 2 (completed) months and 16 years and 1 (completed) month at the beginning of the *testing period*; or as old as between 15 years and 4 (completed) months and 16 years and 3 (completed) months at the beginning of the *testing period*.

**School Level Exclusions** – exclusion of schools from the sampling frame because:

- of geographical inaccessibility (but not part of a region that is omitted from the *PISA Desired Target Population*)
- administration of the PISA assessment within the school would not be feasible
- all students in the school would be *within-school exclusions*
- of other reasons as agreed upon.

**Source Versions** – documents provided in English and French by the International Contractors.

**Target Cluster Size** – the number of students that are to be sampled from schools where not all students are to be included in the sample.

**Testing Period** – the period of time during which data is collected in an *adjudicated entity*.

**Translation Plan** – documentation of all the processes that are intended to be used for all activities related to translation and languages.

**Within-school exclusions** – exclusion of students from potential assessment because of one of the following:

- They are functionally disabled in such a way that they cannot take the PISA test. Functionally disabled students are those with a moderate to severe permanent physical disability.
- They have a cognitive, behavioural or emotional disability confirmed by qualified staff, meaning they cannot take the PISA test. These are students who are cognitively, behaviourally or emotionally unable to follow even the general instructions of the assessment.
- They have insufficient assessment language experience to take the PISA test. Students who have insufficient assessment language experience are those who meet all the following three criteria:
  - they are not native speakers of the assessment language
  - they have limited proficiency in the assessment language
  - they have received less than one year of instruction in the assessment language.
- There are no materials available in the language in which the student is taught.
- They cannot be assessed for some other reason as agreed upon.

## ANNEX G – COMMON AND UNIQUE ITEM PARAMETERS IN EACH DOMAIN, BY COUNTRIES AND LANGUAGES

All tables in Annex G are available online at: [www.oecd.org/pisa](http://www.oecd.org/pisa)



## ANNEX H – SCALAR OR METRIC INVARIANT TREND ITEMS IN EACH DOMAIN

All tables in Annex H are available on line at: [www.oecd.org/pisa](http://www.oecd.org/pisa)



## ANNEX I – PISA CONTRACTORS, STAFF AND CONSULTANTS

PISA is a collaborative effort, bringing together experts from participating countries and economies, steered jointly by their governments on the basis of shared, policy-driven interests.

Each country is represented by members of the PISA Governing Board who determine the policy priorities for PISA, in the context of OECD objectives, and oversee adherence to these priorities during the implementation of the programme. This includes setting priorities for the development of indicators, for establishing the assessment instruments, and for reporting the results.

Experts from participating countries also serve on working groups that are charged with linking policy objectives with the best internationally available technical expertise. By participating in these expert groups, countries ensure that the instruments are internationally valid and take into account the cultural and educational contexts in OECD member and partner countries and economies, that the assessment materials have strong measurement properties, and that the instruments place emphasis on authenticity and educational validity.

Through National Project Managers, participating countries and economies implement PISA at the national level subject to the agreed administration procedures. National Project Managers play a vital role in ensuring that the implementation of the survey is of high quality, and verify and evaluate the survey results, analyses, reports and publications.

The design and implementation of the surveys, within the framework established by the PISA Governing Board, is the responsibility of external contractors. For PISA 2015, the overall management of contractors and implementation was carried out Educational Testing Service in the United States as the Core 7 contractor. The additional tasks related to the implementation of PISA 2015 were implemented through six additional contractors – Cores 1 to 6.

The development of the cognitive assessments was carried out by Pearson in the United Kingdom as the Core 1 contractor.

Core 2 was led by Educational Testing Service and focused on the development of the computer platform in cooperation with the Centre de Recherche Public Henri Tudor (CRP-HT) in Luxembourg.

Core 3 focused on the instrument development, scaling and analysis and was led by the Educational Testing Service, with cooperation from cApStAn Linguistic Quality Control in Belgium for linguistic quality control, the University of Luxembourg, University of Heidelberg, GESIS and the Center for Educational Technology in Israel for test development, the Unité d'analyse des systèmes et des pratiques d'enseignement (aSPe) at the University of Liège in Belgium for coding training for open-constructed items, the International Association for Evaluation of Educational Achievement (IEA) in the Netherlands for the data management software, and HallStat SPRL in Belgium for translation referee.

Core 4 focused on Survey Operations and was implemented by Westat in the United States.

Core 5 focused on sampling and was implemented by Westat in the United States in cooperation with the Australian Council for Educational Research (ACER) for the sampling software KeyQuest.

Core 6 focused on the questionnaire frameworks and instrument development and was carried out by the Deutsches Institut für Internationale Pädagogische Forschung (DIPF) in Germany, with cooperation from Statistics Canada.

The OECD Secretariat has overall managerial responsibility for the programme, monitors its implementation daily, acts as the secretariat for the PISA Governing Board, builds consensus among countries and serves as the interlocutor between the PISA Governing Board and the international Consortium charged with implementing the activities. The OECD Secretariat also produces the indicators and analyses and prepares the international reports and publications in co-operation with the PISA Consortium and in close consultation with member and partner countries and economies both at the policy level (PISA Governing Board) and at the level of implementation (National Project Managers).

**PISA Governing Board** (\* Former PGB member who was involved in PISA 2015)  
Chair of the PISA Governing Board: Michelle Bruniges and Lorna Bertrand\*

### **OECD countries and Associates**

**Australia:** Rhyan Bloor, Michelle Bruniges and Tony Zanderigo\*

**Austria:** Mark Német

**Belgium:** Isabelle Eraud, Geneviève Hindryckx and Christiane Blondin\*



**Brazil:** Maria Helena Guamaaes Castro, Maria Inês Fini, and Luiz Claudio Costa\*

**Canada:** Tomasz Gluszynski, Kathryn O'Grady, Pierre Brochu\* and Patrick Bussiere\*

**Chile:** Carolina Flores, Claudia Matus and Leonor Cariola Huerta\*

**Czech Republic:** Tomas Zatloukal and Jana Paleckova\*

**Denmark:** Mette Hansen, Frida Poulsen, Elsebeth Aller\* and Tine Bak\*

**Estonia:** Maie Kitsing

**Finland:** Tommi Karjalainen

**France:** Thierry Rocher and Bruno Trosseille\*

**Germany:** Martina Diedrich, Katharina Koufen, Elfriede Ohmberger, Annemarie Klemm\* and Susanne von Below\*

**Greece:** Chryssa Sofianopoulou and Vassilia Hatzinikita\*

**Hungary:** Sándor Brassói and Benő Csapó\*

**Iceland:** Stefán Baldursson and Júlíus Björnsson\*

**Ireland:** Peter Archer, Jude Cosgrove\* and Gerry Shiel\*

**Israel:** Hagit Glickman and Michal Beller\*

**Italy:** Roberto Ricci and Paolo Sestito\*

**Japan:** Akiko Ono, Masaharu Shiozaki and Ryo Watanabe\*

**Korea:** Bu Ho Nam, Jimin Cho, Jea Yun Park\*, Sungsook Kim\*, Keunwoo Lee\* and Myungae Lee\*

**Latvia:** Andris Kangro, Aljona Babiča, Ennata Kivrina\* and Dita Traida\*

**Luxembourg:** Amina Kafai

**Mexico:** Eduardo Backhoff Escudero, Ana María Acevess Estrada, Otto Granados Roldán and Francisco Ciscomani\*

**Netherlands:** Marjan Zandbergen and Paul van Oijen\*

**New Zealand:** Craig Jones, Lisa Rodgers\* and Lynne Whitney\*

**Norway:** Marthe Akselsen, Anne-Berit Kavli\* and Alette Schreiner\*

**Poland:** Piotr Mikiewicz, Jerzy Wisniewski\*, Hania Bouacid\* and Stanislaw Drzazdzewski\*

**Portugal:** Hélder Manuel Diniz de Sousa, Luisa Canto\* and Castro Loura\*

**Slovak Republic:** Romana Kanovska and Paulina Korsnakova\*

**Slovenia:** Andreja Barle Lakota, Mojca Straus and Ksenija Bregar-Golobic

**Spain:** Carmen Tovar Sanchez, Vicente Alcañiz Miñano\* and Ismael Sanz Labrador\*

**Sweden:** Eva Lundgren and Anita Wester\*

**Switzerland:** Vera Husfeldt and Claudia Zahner Rossier

**Turkey:** Kemal Bulbul, Mustafa Nadir Çalis\* and Nurcan Devici\*

**United Kingdom:** Lorna Bertrand and Jonathan Wright

**United States:** Peggy Carr, Dana Kelly\*, Jack Buckley\* and Daniel McGrath\*

### **Observers (Partner economies)**

**Albania:** Zamira Gjini and Ermal Elezi\*

**Algeria:** Samia Mezaib and Mohamed Chaibeddra Tani\*

**Argentina:** Elena Duro, Martín Guillermo Scasso\* and Liliana Pascual\*

**Azerbaijan (Baku City only):** Emin Amrullayev

**Belarus (Republic of):** Aliaksandr Yakabchuk and Mikalai Fiaskou

**Bosnia and Herzegovina:** Maja Stojkic

**Brunei Darussalam:** Dr. Azman Ahmad

**Bulgaria:** Neda Kristanova

**Beijing-Shanghai-Jiangsu-Guangdong (China):** Jun Fang, Shiliang Lin and Ping Luo\*

**Colombia:** Ximena Dueñas and Adriana Molina\*

**Costa Rica:** Alicia Vargas and Leonardo Garnier Rimolo\*

**Croatia:** Michelle Bras Roth

**Dominican Republic:** Ancell Scheker Mendoza

**Former Yugoslav Republic of Macedonia:** Natasha Janevska (PISA 2018) and Dejan Zlatkovski\*

**Georgia:** Tamar Bregadze and Natia Mzhavanadze\*

**Hong Kong (China):** Ho-pun Choi, Fanny Yuen-fan Wan and Esther Sui-chu Ho\*

**Indonesia:** Dr. Totok Suprayitno, Furqon Furqon\* and Khairil Anwar Notodiputro\*

**Jordan:** Khattab Mohammad Abulibdeh

**Kazakhstan:** Shamshieva Nurgul, Serik Irsaliyev\* and Almagul Kultumanova\*

**Kosovo:** Anila Statovci Demaj

**Lebanon:** Nada Ouweijan

**Lithuania:** Rita Dukynaite

**Macao (China):** Leong Lai

**Malaysia:** Hon. Dato' Sulaiman bin Wak, Khairil Awang\* and Amin Senin\*

**Malta:** Charles Mifsud

**Moldova (Republic of):** Anatolie Topala and Valeriu Gutu\*

**Montenegro:** Dragana Dmitrovic and Zeljko Jacimovic\*

**Morocco:** Mohammed Sassi

**Panama:** Marelissa Tribaldos

**Peru:** Humberto Hildebrando Pérez León Ibañez and Liliana Miranda Molina\*

**Philippines:** Elvin Ivan Yaw



**Qatar:** Khalid Abdulla Al-Harqan and Hamda Al Sulaiti\*

**Romania:** Roxana Mihail and Daniela Bogdan

**Russian Federation:** Galina Kovaleva, Sergey Kravtsov and Isak Froumin\*

**Saudi Arabia:** Mohamed Al-harthi

**Serbia (Republic of):** Anamarija Viček and Zorana Lužanin\*

**Singapore:** Chern Wei Sng and Khah Gek Low\*

**Chinese Taipei:** Tian-Ming Sheu, Peng Li-Chun\*, Gwo-Dong Chen\* and Chih-Wei Hue\*

**Thailand:** Supattra Pativisan and Precharn Dechsri\*

**Trinidad and Tobago:** Mervyn Sambucharan and Harrilal Seecharan

**Tunisia:** Riadh Ben Boubaker

**Ukraine:** Pavlo Khobzey

**United Arab Emirates:** Hessa Alwahabi, Rabaa Alsumaiti, Moza al Ghufly\*, Ayesha G. Khalfan Almerri\*, Ali Jaber Al Yafei\* and Khawla Al Mualla\*

**Uruguay:** Andrés Peri and María Helvecia Sanchez Nunez\*

**Viet Nam:** Le Thi My Ha

### PISA 2015 National Project Managers (\* Former PISA 2015 NPM)

**Albania:** Rezana Vrapi and Alfons Harizaj\*

**Algeria:** Samia Mezaib

**Argentina:** Liliana Pascual

**Australia:** Sue Thomson

**Austria:** Birgit Suchan

**Beijing-Shanghai-Jiangsu-Guangdong (China):** Wang Lei

**Belgium:** Inge De Meyer and Anne Matoul

**Brazil:** Aline Mara Fernandes

**Bulgaria:** Svetla Petrova

**Canada:** Pierre Brochu and Tamara Knighton\*

**Chile:** Ema Lagos Campos

**Colombia:** Javier Juyar, Francisco Reyes\*, Adriana Molina\* and Julián P. Mariño\*

**Costa Rica:** Lilliam Mora

**Croatia:** Michelle Bras Roth

**Czech Republic:** Radek Blažek and Jana Palecková\*

**Denmark:** Hans Hummelgaard, Niels Egelund\* and Chantal Nielsen\*

**Dominican Republic:** Massiel Cohen

**Estonia:** Gunda Tire

**Finland:** Jouni Välijärvi

**Former Yugoslav Republic of Macedonia:** Natasha Janevska and Dejan Zlatkovski

**France:** Irène Verlet

**Georgia:** Natia Mzhavanadze

**Germany:** Christine Sälzer and Manfred Prenzel

**Greece:** Chryssa Sofianopoulou

**Hong Kong (China):** Esther Sui-chu Ho

**Hungary:** László Ostorics

**Iceland:** Almar Midvik Halldorsson

**Indonesia:** Ir. Nizam

**Ireland:** Gerry Shiel

**Israel:** Joel Rapp and Inbal Ron-Kaplan

**Italy:** Carlo Di Ciacchio

**Japan:** Akiko Ono

**Jordan:** Emad Ababneh

**Kazakhstan:** Irina Imanbek and Gulmira Berdibayeva\*

**Korea:** Jaok Ku, Jimin Cho\* and Mi-Young Song\*

**Latvia:** Andris Kangro

**Lebanon:** Bassem Issa and Antoine Skaf\*

**Lithuania:** Mindaugas Stundza

**Luxembourg:** Bettina Boehm

**Macao (China):** Kwok Cheung Cheung

**Malaysia:** Muhammad Zaini Mohd Zain

**Malta:** Louis Scerri

**Mexico:** María Antonieta Díaz Gutierrez

**Moldova (Republic of):** Valeriu Gutu

**Montenegro:** Divna Paljevic Sturm

**Netherlands:** Jesse Koops and Johanna Kordes\*

**New Zealand:** Steve May, Saila Cowles and Maree Telford\*

**Norway:** Marit Kjaernsli



**Peru:** Liliana Miranda Molina  
**Poland:** Barbara Ostrowska  
**Portugal:** João Maroco  
**Qatar:** Shaikha Al-Ishaq and Saada Al-Obaidli\*  
**Romania:** Silviu Cristian Mirescu  
**Russian Federation:** Galina Kovaleva  
**Serbia:** Dragica Pavlovic-Babic  
**Singapore:** Chew Leng Poon, Elaine Chua and Pik Yen Lim\*  
**Slovak Republic:** Jana Ferencova  
**Slovenia:** Mojca Straus  
**Spain:** Lis Cercadillo Pérez  
**Sweden:** Magnus Oskarsson  
**Switzerland:** Christian Nidegger  
**Chinese Taipei:** Hsiao-Ching She and Huann-Shyang Lin  
**Thailand:** Nantawan Nantawanit and Suchada Thaithae  
**Trinidad and Tobago:** Mervyn Sambucharan  
**Tunisia:** Mehrez Drissi and Med Kamel Essid\*  
**Turkey:** Umut Erkin Taş  
**United Arab Emirates:** Mouza Rashed Khalfan Al Ghufli  
**United Kingdom:** Dawn Pollard and Juliet Sizmur  
**United States:** Dana Kelly, Patrick Gonzales and Holly Xie\*  
**Uruguay:** Maria Helvacia Sánchez Nunez  
**Viet Nam:** Thi My Ha Le

### **OECD Secretariat**

Andreas Schleicher (Strategic development)  
Marilyn Achiron (Editorial support)  
Peter Adams (Project management)  
Francesco Avvisati (Analytic services)  
Yuri Belfali (Strategic development)  
Rose Bolognini (Editorial and production support)  
Guillaume Bousquet (Analytic services)  
Jenny Bradshaw (Project management 2015)  
Esther Carvalhaes (Analytic services)  
Claire Chetcuti (Administrative support)  
Anna Choi (Analytic services)  
Cassandra Davis (Dissemination co-ordination)  
Alfonso Echazarra (Analytic services)  
Juliet Evans (Administration and partner country/economy relations)  
Hélène Guillou (Analytic services)  
Carlos González-Sancho (Analytic services)  
Tue Halgreen (Project management)  
Miyako Ikeda (Analytic services)  
Thomas Marwood (Administrative support)  
Jeffrey Mo (Analytic services)  
Chiara Monticone (Analytic services)  
Lesley O'Sullivan (Administrative support)  
Bonaventura Francesco Pacileo (Analytic services)  
Judit Pál (Analytic services)  
Mario Piacentini, (Analytic services)  
Giannina Rech (Analytic services)  
Daniel Salinas (Analytic services)  
Michael Stevenson (Dissemination co-ordination)  
Hanna Varkki (Administrative support)  
Sophie Vayssettes (Project management)

### **PISA 2015 science expert group**

Jonathan Osborne (SEG Chair) (Stanford University, United States and United Kingdom)  
Marcus Hammann (Munster University, Germany)  
Sarah Howie (University of Pretoria, South Africa)  
Jody Clarke-Midura (Harvard University, United States)  
Robin Millar (University of York, United Kingdom)  
Andrée Tibergien (University of Lyon, France)  
Russell Tytler (Deakin University, Australia)  
Darren Wong (National Institute of Education, Singapore)



### **Extended group**

Rodger Bybee (Biological Sciences Curriculum Study (BSCS), United States)  
 Jens Dolin (University of Copenhagen, Denmark)  
 Harrie Eijkelhof (Utrecht University, Netherlands)  
 Geneva Haertel (SRI, United States)  
 Michaela Mayer (University of Roma Tre., Italy)  
 Eric Snow (SRI, United States)  
 Manabu Sumida (Ehime University, Japan)  
 Benny Yung (University of Hong Kong, Hong Kong, China)

### **PISA 2015 problem solving expert group**

Arthur Graesser (Chair) (The University of Memphis United States)  
 Eduardo Cascallar (Katholieke Universiteit Leuven, Belgium)  
 Pierre Dillenbourg (Ecole Polytechnique Fédérale de Lausanne, Switzerland)  
 Patrick Griffin (University of Melbourne, Australia)  
 Chee Kit Looi (Nanyang Technological University, Singapore)  
 Jean-François Rouet (University of Poitiers, France)

### **Extended group**

Rafael Calvo (University of Sydney, Argentina)  
 Tak Wai Chan (National Central University of Taiwan, China)  
 Stephen Fiore (University of Central Florida, USA)  
 Joachim Funke (University of Heidelberg, Germany)  
 Manu Kapur (National Institute of Education, Singapore)  
 Naomi Miyake (University of Tokyo, Japan)  
 Yigal Rosen (University of Haifa, Israel)  
 Jennifer Wiley (University of Illinois at Chicago, USA)

### **PISA 2015 questionnaire expert group**

David Kaplan (Chair as of 2014) (University of Wisconsin-Madison, United States)  
 Eckhard Klieme (Chair until 2013) (German Institute for International Educational Research, Germany (DIPF), Frankfurt, Germany)  
 Gregory Elacqua (Universidad Diego Portales, Chile)  
 Marit Kjærnsli (University of Oslo, Norway)  
 Leonidas Kyriakides (University of Cyprus, Cyprus)  
 Henry M. Levin (Columbia University, United States)  
 Naomi Miyake (University of Tokyo, Japan)  
 Jonathan Osborne (Stanford University, United States)  
 Kathleen Scalise (University of Oregon, United States)  
 Fons van de Vijver (Tilburg University, Netherlands)  
 Ludger Wößmann (University of Munich, Germany)

### **Technical advisory group**

Keith Rust (chair) (Westat, USA)  
 Theo Eggen (Cito, Netherlands)  
 John de Jong (Pearson, UK/VU University Amsterdam, Netherlands)  
 Jean Dumais (Statistics Canada, Canada)  
 Cees Glas (University of Twente, Netherlands)  
 David Kaplan (University of Wisconsin-Madison, USA and DIPF, Germany)  
 Irwin Kirsch (ETS, USA)  
 Christian Monseur (Université de Liège, Belgium)  
 Sophia Rabe-Hesketh (University of Berkeley, USA)  
 Thierry Rocher (Ministère de l'Éducation Nationale, France)  
 Leslie A. Rutkowski (University of Oslo, Norway)  
 Margaret Wu (Victoria University, Australia)  
 Kentaro Yamamoto (ETS, USA)

### **PISA 2015 Lead Contractors**

#### **Educational Testing Service (United States) – Cores 2, 3 and 7 lead contractor**

Irwin Kirsch (International project director)  
 Claudia Tamassia (International project manager)  
 David Garber (Project management, paper booklets and coding)  
 Larry Hanover (Editorial support)  
 Lisa Hemat (Project support)  
 Isabelle Jars (Project management, questionnaires)  
 Judy Mendez (Project support and contracts)



Eugenio Gonzalez (Training and data products)  
 Kentaro Yamamoto (Director, psychometrics and analysis)  
 Matthias von Davier (Director, psychometrics and analysis)  
 Chentong Chen (Psychometrics and analysis)  
 Haiwen Chen (Psychometrics and analysis)  
 Qiwei He (Psychometrics and analysis)  
 Lale Khorramdel (Manager, psychometrics and analysis)  
 Hyo Jeong Shin (Psychometrics and analysis)  
 Jon Weeks (Psychometrics and analysis)  
 Marylou Lennon (Test development coordinator, science and collaborative problem solving)  
 Eric Steinhauer (Test Development, Lead, Science and Collaborative Problem Solving)  
 Janet Koster van Groos (Test Development, Science)  
 Marshall L Freedman (Test Development Science)  
 Israel Solon (Test Development Science)  
 Jakub Novak (Test Development Science)  
 Nancy Olds (Test Development Science)  
 Paul Borysewicz (Test Development, Collaborative Problem Solving)  
 William Sims (Test Development, Collaborative Problem Solving)  
 Peter Cooper (Test Development, Collaborative Problem Solving)  
 Michael Wagner (Director, platform development)  
 Jason Bonthon (Platform development and authoring)  
 Paul Brost (Platform development)  
 Ramin Hemat (Platform development and authoring)  
 Keith Keiser (Platform development and coding system)  
 Debbie Pisacreta (Interface design and graphics)  
 Janet Stumper (Graphics)  
 Ted Blew (Director, data analysis, research and technology)  
 John Barone (Director, data analysis and database preparation)  
 Mathew Kandathil (Leader, data analysis and data management)  
 Kevin Bentley (Data products)  
 Hezekiah Bunde (Data management)  
 Karen Castellano (Data analysis)  
 Scott Davis (Data analysis)  
 Chantal Delaney (Data management)  
 Matthew Duchnowski (Data management)  
 Ying Feng (Data management)  
 Zhumei Guo (Data analysis)  
 Laura Jerry (Data analysis)  
 Lokesh Kapur (Data analysis)  
 Debra Kline (Data analysis leader)  
 Phillip Leung (Data products leader)  
 Alfred Rogers (Data management leader)  
 Carla Tarsitano (Data management leader)  
 Sarah Venema (Data products)  
 Tao Wang (Data products)  
 Lingjun Wong (Data analysis)  
 Yan Zhang (Data management)  
 Wei Zhao (Data analysis)

#### **Deutsches Institut für Internationale Pädagogische Forschung<sup>1</sup> (DIPF, GERMANY) – Core 6 lead contractor**

Eckhard Klieme (Study director, questionnaire framework and development)  
 Nina Jude (Management and questionnaire development)  
 Sonja Bayer (Questionnaire development and analysis)  
 Janine Buchholz (Questionnaire scaling)  
 Frank Goldhammer (Questionnaire development)  
 Silke Hertel (Questionnaire development)  
 Franz Klingebiel (Questionnaire development)  
 Susanne Kuger (Questionnaire framework and development)  
 Ingrid Mader (Team assistance)  
 Tamara Marksteiner (Questionnaire analysis)  
 Jean-Paul Reeff (International Consultant)  
 Nina Roczen (Questionnaire development)  
 Brigitte Steinert (Questionnaire development)  
 Svenja Vieluf (Questionnaire development)

1. Also referred to as the German Institute for International Educational Research



### **Pearson (UNITED KINGDOM) – Core 1 lead contractor**

John de Jong (Programme director)  
 Catherine Hayes (Programme manager)  
 Elise Bromley (Programme administrator)  
 Rose Clesham (Content lead, scientific literacy)  
 Peter Foltz (Content lead, collaborative problem solving)  
 Christine Rozunick (Content lead, scientific literacy)  
 Jon Twing (Psychometric consultant)  
 Michael Young (Psychometric consultant)

### **WESTAT (UNITED STATES) – Cores 4 and 5 lead contractor**

Keith Rust (Director of the PISA consortium for sampling and weighting)  
 Sheila Krawchuk (Sampling, weighting and quality monitoring)  
 Andrew Caporaso (Weighting)  
 Jessica Chan (Sampling and weighting)  
 William Chan (Weighting)  
 Susan Fuss (Sampling and weighting)  
 Amita Gopinath (Sampling and weighting)  
 Evan Gutentag (Weighting)  
 Jing Kang (Sampling and weighting)  
 Veronique Lieber (Sampling and weighting)  
 John Lopdell (Sampling and weighting)  
 Shawn Lu (Weighting)  
 Martha Rozsi (Weighting)  
 Yumiko Siegfried (Sampling and weighting)  
 Joel Wakesberg (Sampling and weighting)  
 Sipeng Wang (Weighting)  
 Erin Wiley (Sampling and weighting)  
 Sergey Yagodin (Weighting)  
 Merl Robinson (Director of Core 4 Contractor for Survey Operations)  
 Michael Lemay (Manager of Core 4 Contractor for Survey Operations)  
 Jessica Chan (National Centre Support, Quality Control)  
 Lillian Diaz-Hoffman (National Centre Support, Quality Control)  
 Sarah Hartge (National Centre Support, Quality Control)  
 Beverley McGaughan (National Centre Support, Quality Control)

### **PISA 2015 Contributors, working with Lead Contractors**

#### **Australian Council for Educational Research (AUSTRALIA) – Core 5 contributor**

Eveline Gebhardt (Project director)  
 Alla Routitsky (Within-school sampling)  
 Charlotte Waters (Within-school sampling)  
 Jorge Fallas (Within-school sampling)  
 Renee Chow (Within-school sampling)  
 David Tran (Programmer)  
 Martin Murphy (School sampling)  
 Clare Ozolins (School sampling)  
 Greg Macaskill (School sampling)  
 Jennifer Hong (School sampling)  
 Jorge Fallas (School sampling)  
 Renee Chow (School sampling)  
 Thomas Stephen (School sampling)

#### **Center for Educational Technology – Core 3 contributor on test development**

Tali Freund (Test Development Coordinator, Science and Collaborative Problem Solving)  
 Rachel Mintz (Test Development, Lead, Science)  
 Nurit Keinan (Test Development, Science)  
 Hava Ben-Horin (Test Development, Science)  
 Sherman Rosenfeld (Test Development, Science)  
 Lilach Tencer-Hershkovitz (Test Development, Science)  
 Nadav Caspi (Test Development, Science)  
 Elinor Shaked-Blazer (Test Development, Science)  
 Sara Hershkovitz (Test Development, Lead, Collaborative Problem Solving)  
 Cecilia Waisman (Test Development, Collaborative Problem Solving)  
 Helit Heffer (Test Development, Collaborative Problem Solving)  
 Estela Melamed (Test Development, Science and Collaborative Problem Solving)



### **cApStAn Linguistic Quality Control (BELGIUM) – Core 3 contributor on linguistic quality control**

Steve Dept (Project director, translatability assessment,  
Lieve Deckx (Verification management, cognitive units)  
Andrea Ferrari (Linguistic quality assurance and quality control designs)  
Musab Hayatli (Right-to-left scripts, cultural adaptations)  
Elica Krajceva (Verification management, questionnaires)  
Shinoh Lee (Verification management, cognitive units)  
Irene Liberati (Verification management, cognitive units)  
Roberta Lizzi (Verification management, trend content)  
Laura Wayrynen (Translation and verification operations)

### **GESIS-Leibniz Institute for the Social Sciences (GERMANY) – Core 3 contributor on test development**

Anouk Zabal (Test Development Coordinator, Science and Collaborative Problem Solving, Software Testing)  
Dorothee Behr (Test Development, Science and Collaborative Problem Solving, Software Testing)  
Daniela Ackermann (Test Development, Science and Collaborative Problem Solving, Software Testing)

### **HallStat SPRL (BELGIUM) – Core 3 contributor as the translation referee**

Beatrice Halleux (Consultant, translation/verification referee, French source development)

### **Luxembourg Institute for Science and Technology (LUXEMBOURG) – Core 2 Contributor on the development of the computer-based platform for the background questionnaire and cognitive assessment**

Jehan Bihim (Questionnaire development)  
Joël Billard (Multilingual framework and questionnaire development)  
Cyril Hazotte (System administration)  
Anne Hendrick (Platform Leader, project co-ordination)  
Raynald Jadoul (Project management and software architecture)  
Isabelle Jars (Project management and testing)  
Lionel Lecaque (Software quality and knowledge base administration)  
Primaël Lorbat (Multilingual framework and questionnaire architecture)  
Matteo Melis (Portal integration and questionnaire development)  
Jean-François Merche (System integration and administration)  
Vincent Porro (Lead designer and staff co-ordination)  
Igor Ribassin (Workflow development and offline tools development)  
Somsack Sipasseuth (Workflow development and knowledge base integration)  
Nicolas Yodi (Portal integration and questionnaire development)

### **Statistics Canada (CANADA) – Core 6 contributor on questionnaires**

Sylvie Grenier (Overall management)  
Tamara Knighton (Overall management)  
Isabelle Thorny (Implementation Delivery System)  
Ginette Grégoire (Implementation Delivery System)  
Martine Lafrenière (Implementation Delivery System)  
Rosa Tatasciore (Implementation Delivery System)

### **Unité d'analyse des Systèmes et des Pratiques d'enseignement (aSPe, BELGIUM) – Core 3 contributor on coding training**

Dominique LaFontaine (Project supervisor)  
Ariane Baye (Coding training, reading)  
Isabelle Demonty (Coding training, mathematics)  
Annick Fagnant (Coding training, mathematics)  
Geneviève Hindryckx (Coding training, science)  
Anne Matoul (Coding training, reading)  
Valérie Quittre (Coding training, science)

### **University of Heidelberg (GERMANY) – Core 3 contributor on test development**

Daniel Holt (Test Development, Collaborative Problem Solving)  
Andreas Fischer (Test Development, Collaborative Problem Solving)  
Ursula Pöll (Test Development, Collaborative Problem Solving)  
Julia Hilse (Test Development, Collaborative Problem Solving)  
Saskia Kraft (Test Development, Collaborative Problem Solving)  
Florian Hofmann (Test Development, Collaborative Problem Solving)

### **University of Luxembourg (LUXEMBOURG) – Core 3 contributor on test development**

Romain Martin (Test Development Coordinator, Science)  
Samuel Greiff (Test Development Coordinator, Collaborative Problem Solving)  
Sara Wilmes (Test Development, Science)  
Sophie Doublet (User Testing)  
Vincent Koenig (User Testing)  
Katja Weinerth (User Testing)



## **ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT**

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

OECD Publishing disseminates widely the results of the Organisation's statistics gathering and research on economic, social and environmental issues, as well as the conventions, guidelines and standards agreed by its members.

# PISA 2015 Technical Report

The *PISA 2015 Technical Report* describes the methodology underlying the PISA 2015 survey which tested 15-year-olds' competencies in science, reading and mathematics, and, for some countries, financial literacy and collaborative problem solving. It examines additional features related to the implementation of the project at a level of detail that allows researchers to understand and replicate its analyses. The reader will find a wealth of information on the test and sample design, modes of administration (paper-based or computer-based), methodologies used to analyse the data, technical features of the project, and quality control mechanisms.

## Contents

- Chapter 1. Programme for International Student Assessment: an overview
- Chapter 2. Test design and test development
- Chapter 3. Context questionnaire development
- Chapter 4. Sample design
- Chapter 5. Translation and verification of the survey material
- Chapter 6. Field operations
- Chapter 7. PISA quality monitoring
- Chapter 8. Survey weighting and the calculation of sampling variance
- Chapter 9. Scaling PISA data
- Chapter 10. Data management procedures
- Chapter 11. Sampling outcomes
- Chapter 12. Scaling outcomes
- Chapter 13. Coding design, coding process, and coder reliability studies
- Chapter 14. Data adjudication
- Chapter 15. Proficiency scale construction
- Chapter 16. Scaling procedures and construct validation of context questionnaire data
- Chapter 17. Questionnaire design and computer-based questionnaire platform
- Chapter 18. Computer-based texts
- Chapter 19. International data products

## THE OECD PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT (PISA)

PISA does not just ascertain whether students can reproduce knowledge; it also examines how well students can extrapolate from what they have learned and can apply that knowledge in unfamiliar settings, both in and outside of school. This approach reflects the fact that modern economies reward individuals not for what they know, but for what they can do with what they know.

PISA's unique features include its:

- policy orientation, which connects data on student learning outcomes with data on students' backgrounds and attitudes towards learning, and on key factors that shape their learning in and outside school, in order to highlight differences in performance patterns and identify the characteristics of schools and education systems that perform well
- innovative concept of "literacy", which refers to students' capacity to apply knowledge and skills in key subjects, and to analyse, reason and communicate effectively as they identify, interpret and solve problems in a variety of situations
- relevance to lifelong learning, as PISA asks students to report on their motivation to learn, their beliefs about themselves and their learning strategies
- regularity, which enables countries to monitor their progress in meeting key learning objectives
- breadth of coverage, which, in PISA 2015, encompasses the 35 OECD countries and 37 partner countries and economies.